# Positive Selektion in langlebigen Nagern und kurzlebigen Fischen – Eine bioinformatische Suche nach der genetischen Basis des Alterns

Dissertation

zur Erlangung des akademischen Grades

„doctor rerum naturalium“ (Dr. rer. nat.)

vorgelegt dem Rat der Biologisch-Pharmazeutischen Fakultät

der Friedrich-Schiller- Universität Jena

von M. Sc. (Bioinformatik und Genomforschung) Arne Sahm

geboren am 01.10.1989 in Königs Wusterhausen

Die vorliegende Arbeit wurde in der Zeit vom September 2013 bis Juni 2017 am Leibniz-Institut für Alternsforschung - Fritz-Lipmann-Institut e.V. (FLI) angefertigt.

Gutachter:

1. PD Dr. Matthias Platzer (Jena)
2. Prof. Dr. Hans Kestler (Ulm)
3. Prof. Dr. Thomas Wiehe (Köln)

Datum der Verteidigung: 27.11.2017

# Inhalt

# Zusammenfassung

Zahlreiche Geschichten und Legenden, angefangen von den frühen Hochkulturen bis hinein in die moderne Populärliteratur, zeugen von dem Wunsch der Menschen das Altern zu verlangsamen. Viele Forscher gehen davon aus, dass genau dies notwendig sein wird, um den seit einigen Jahrzehnten anhaltenden Trend steigender Lebenserwartungen auch in Zukunft fortsetzen zu können und halten es für wahrscheinlich, dass ein verbessertes Verständnis des Alternsprozesses uns dahin führen wird.

Um zu einem solchen Verständnis beizutragen, habe ich in dieser Arbeit mit bioinformatischen Mitteln nach der genetischen Basis natürlicherweise Weise vorhandener Unterschiede in der Lebenserwartung verschiedener Spezies gefahndet. Ich bediente mich dabei der Methode der genomweiten Suche nach positiv selektierten Genen (PSGs), die auf dem Vergleich der proteinkodierenden Sequenzen verschiedener Spezies beruht. Dazu habe ich verschiedene Software-Werkzeuge entwickelt und diese in dem Programm PosiGene zusammengefasst. PosiGene ist das erste öffentliche verfügbare Programm, das es ermöglicht PSGs genomweit auf beliebigen, vom Nutzer ausgewählten evolutionären Zweigen zu detektieren. Das erlaubt es allgemein – also über die in dieser Arbeit im Fokus stehende Alternsforschung hinaus – nach der genetischen Basis speziesspezifischer, phänotypischer Merkmale zu suchen. Die Offenlegung des Quellcodes in GitHub und die Bereitstellung einer installationslos, direkt nach dem Herunterladen einsatzfähigen Software ermöglicht eine breite Anwendung sowie eine kollektive Weiterentwicklung des Programmpakets zum Nutzen der wissenschaftlichen Gemeinschaft (Manuskript I).

Konkret untersuchte ich mittels PosiGene zum einen die Verkürzung von Lebensspannen auf evolutionären Zweigen der Prachtgrundkärpflinge (*Nothobranchius*). Diese zählen mit Lebenserwartungen von z.T. nur sechs Monaten zu den kurzlebigsten Wirbeltierarten überhaupt. Zum anderen untersuchte ich die Verlängerung von Lebensspannen auf Zweigen der Nagetierfamilie der Sandgräber (Bathyergidae), deren Vertreter mit bis zu 30 Jahren um ein Vielfaches älter werden als die meisten anderen Nagetiere.

Die erste Anwendung von PosiGene erfolgte im Rahmen der Genompublikation des Türkisen Prachtgrundkärpflings – des kurzlebigsten Vertreters der Gattung. Auf dem evolutionären Zweig dieser Spezies konnte ich sieben PSGs detektieren, von denen fünf während des Alterns bei diesem Fisch differentiell exprimiert werden und zwei bereits eine Rolle in der Alternsforschung spielen (Manuskript II). Zeitgleich damit erschien eine weitere Genompublikation des Türkisen Prachtgrundkärpflings, in der mehr als 500 PSGs detektiert worden waren. Ich bin der Frage nachgegangen, wie der große Unterschied in der Zahl der identifizierten PSGs – sieben zu 500 – zustande kommen konnte und habe exemplarisch gezeigt, dass die Auswahl der in die Analyse einbezogenen Spezies entscheidend sowohl für die Vergleichbarkeit als auch die Aussagekraft solcher Untersuchungen ist (Manuskript III). Zudem habe ich auf Zweigen von Vorfahren der heute lebenden Prachtgrundkärpflingsarten nach PSGs gefahndet, auf denen die Lebensspanne sehr wahrscheinlich stark verkürzt wurde. Im Ergebnis wurden PSGs in allen Schritten der mitochondrialen Biogenese identifiziert, wobei die meisten der Schritte im Sinne paralleler Evolution auf mehreren Zweigen gleichzeitig betroffen waren. Meine Arbeit unterstützt insbesondere mehrere experimentelle Studien, die vorgeschlagen haben, dass die koordinierte Synthese der Komponenten der mitochondrialen Atmungskette ein entscheidender Faktor für die Länge der Lebensspanne ist (Manuskript IV).

Im Zusammenhang mit der Evolution von Langlebigkeit bei Sandgräbern habe ich festgestellt, dass das Immunsystem, die antioxidantielle Verteidigung, der Regulator von mTOR sowie Prozesse, die ihrerseits von mTOR reguliert werden wie bspw. Translation, Autophagie und – wie schon zuvor bei der Evolution

von Kurzlebigkeit – die mitochondriale Biogenese von positiver Selektion betroffen waren. Analysen der Expressionsrichtungen der PSGs während des Alterns von kurzlebigen Ratten und langlebigen Nacktmullen resultierten darüber hinaus in signifikanten und gegensätzlichen Expressionsmustern, die mit der Alternstheorie der antagonistischen Pleiotropie im Einklang stehen (Manuskript V).

Insgesamt unterstreicht die Arbeit die Relevanz verschiedener biologischer Prozesse und Hypothesen für die Alternsforschung und liefert in diesem Zusammenhang ebenfalls eine Reihe vielversprechender Genkandidaten, die als Ansatzpunkte für funktionelle Folgestudien genutzt werden können. Mit der Veröffentlichung von PosiGene habe ich zugleich die zur Generierung dieser Einsichten verwendete Methode fortentwickelt. Ich sehe meine Arbeit als einen Schritt hin zu einem verbesserten Verständnis des Alterns, das es uns eines, hoffentlich nicht allzu fernen Tages ermöglichen wird die Legenden über die Verlangsamung dieses Prozesses – und damit ein längeres und gesünderes Leben – ein Stück weit Wirklichkeit werden zu lassen.

## Summary

Many stories and legends, from the ancient civilizations to the modern popular literature, give evidence of the human wish to slow down aging. Numerous scientists assume that exactly this will be necessary to continue the lasting trend of rising life expectations from the last decades also in future. In addition, they consider it as a probable scenario that an improved understanding of the aging process will lead us to that point.

To contribute to such an understanding, in this work I search by bioinformatical means for the genetic basis of natural differences in lifespan between species. I use the method of genome-wide searches for positively selected genes (PSGs) which is based on the comparison of protein coding sequences from different species. To this end, I developed several software tools and combined them into the program PosiGene. PosiGene is the first publicly available program that enables genome-wide detection of PSGs on arbitrary, user-chosen evolutionary branches. This allows to search in a general way – i.e. beyond aging research which is the focus of this work – to search for the genetic basis of species-specific phenotypic traits (Manuscript I). The publication of the source code in GitHub and the preparation of a software that is operational directly after download enables a broad application and a collective, further development of the program package for the benefit of the scientific community.

In particular, I examine on one hand the reduction of lifespans on evolutionary branches of the *Nothobranchius*. Species from this fish genus are among the shortest-lived known vertebrates – with life expectancies that are in part only six months. On the other hand, I examine the extension of lifespans on branches of the rodent family of African mole-rats (Bathyergidae) whose representatives reach lifespans that are multiple times as long as those of most other rodent species.

PosiGene was first applied in the context of the genome publication of *Nothobranchius furzeri* – the shortest-lived representative of the genus. On the evolutionary branch of this species I detected seven genes under positive selection. Five of these genes were differentially expressed during aging and two are already known to play a role in aging research (Manuscript II). At the same time, another genome publication of *N. furzeri* was published that detected more than 500 PSGs. I examined the question of how the big difference in the number of identified PSGs – 500 vs. seven – could emerge and showed that the selection of species that are included in the analysis is decisive both for comparability and explanatory power of such studies (Manuscript III). Furthermore, I searched for PSGs on branches of ancestors of today living *Nothobranchius* species on which lifespan was probably strongly reduced. As a result, PSGs in all steps of mitochondrial biogenesis were identified. Moreover, most of these steps were affected by positive selection on multiple branches in the sense of parallel evolution. My work supports especially multiple experimental studies that suggested that the coordinated synthesis of components of the mitochondrial respiratory chain is an important factor for the length of lifespan (Manuscript IV).

In the context of evolution of longevity in African mole-rats I found the immune system, the antioxidant defense, the regulator of mTOR as well as processes that are regulated by mTOR such as translation, autophagy and – as seen before for the evolution of short lifespans – mitochondrial biogenesis were affected by positive selection. Analyses of expression directions of PSGs during aging of short-living rats and long-living naked mole-rats resulted in significant and contrary expression patterns that fit the aging theory of antagonistic pleiotropy (Manuscript V).

Altogether, my work emphasizes the relevance of multiple biological processes and hypotheses for aging research and provides in this context a series of promising gene candidates that can be used as starting points for functional follow-up studies. At the same time, by the publication of PosiGene I improved the method that was used to gain the mentioned insights. I see my work as a step towards an improved

understanding of aging that will hopefully allow us to make the legends of slowed down aging become true to some extent in a not too distant future – and with that a longer and healthier life.

.

# Einleitung

## 1. Altern im historischen und gesellschaftlichen Kontext

Schon seit jeher waren Menschen von der Vorstellung fasziniert das Altern verlangsamen, anhalten oder sogar umkehren zu können. So sucht und findet Gilgamesch im gleichnamigen, aus der frühen Hochkultur Mesopotamiens stammenden Epos eine Pflanze, die alten Menschen die Jugend zurückbringt. In der griechischen Mythologie sorgt Ambrosia, die Speise der Götter, dafür, dass diese und ihre sterblichen Günstlinge niemals altern. Einen ähnlichen Effekt haben in der nordischen Mythologie die Äpfel der Göttin Idun, in den abendländischen Sagen der heilige Gral und in der modernen Populärliteratur, um nur eines von vielen Beispielen zu nennen, die Ringe der Macht in den Werken J.R.R. Tolkiens.

Tatsächlich ist es insbesondere den modernen Industriegesellschaften gelungen, die Lebenserwartung in den letzten anderthalb Jahrhunderten stark zu erhöhen, sodass sie dort gegenwärtig so hoch ist wie nie zuvor in der Menschheitsgeschichte. In Deutschland stieg sie von weniger als 40 Jahren im Erhebungszeitraum von 1871-1881 auf etwa 80 Jahre im neuen Jahrtausend (Statistisches Bundesamt 2016) (Abbildung 1). Allerdings werden die Steigerungen im späten 19. und frühen 20. Jahrhundert vorwiegend auf eine verbesserte Kontrolle der Verbreitung infektiöser Krankheiten, die Verringerung der Kindersterblichkeit und Verbesserungerungen des Lebensstandards der weniger begüterten sozialen Schichten zurückgeführt (Wilmoth 2000; Oeppen and Vaupel 2002; Sierra, et al. 2009; Olshansky 2015). Damit wurden zunächst vor allem die Todesraten von Jüngeren massiv gesenkt. Erst ab etwa Mitte des 20. Jahrhunderts gilt die Verminderung der Todesraten bei älteren Personen als Hauptfaktor für die Erhöhung der Lebenserwartung (Wilmoth 2000; Oeppen and Vaupel 2002; Christensen, et al. 2009; Olshansky 2015). Parallel zur durchschnittlichen Lebensspanne wurde auch die maximale Lebensspanne erhöht: In Schweden von durchschnittlich 101 Jahren in den 1860ern auf 108 Jahren in den 1990ern (Wilmoth, et al. 2000).

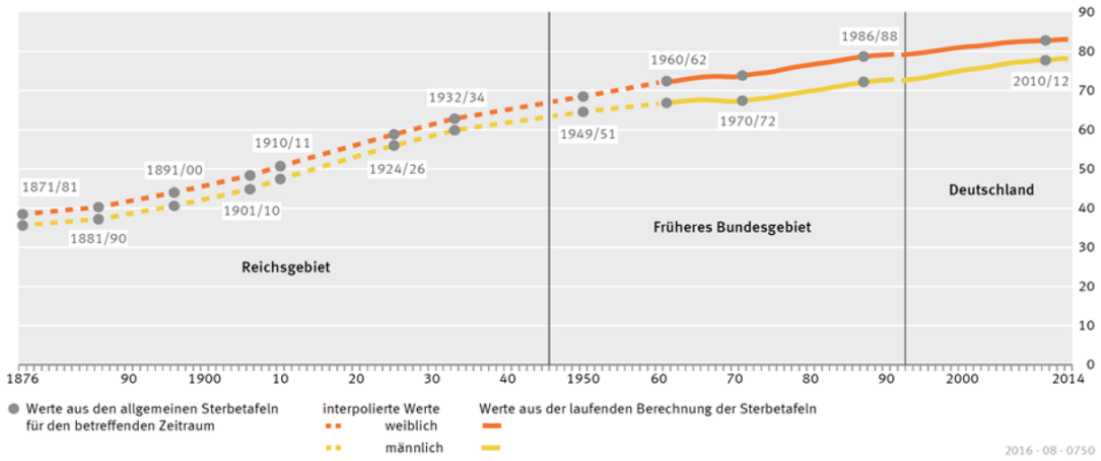



**Abbildung 1.** Entwicklung der Lebenserwartung Neugeborener in Deutschland seit 1871/1881. Entnommen aus: (Statistisches Bundesamt 2016).

Die wenigsten Menschen dürften sich jedoch eine Erhöhung ihrer Lebenspanne wünschen, die sich vor allem aus einer Verlängerung des Siechtums speist. Die Gefahr eines längeren Siechtums ist unmittelbar dadurch gegeben, dass für viele Krankheiten, insbesondere solche die die Lebensqualität stark vermindern können, die Erkrankungswahrscheinlichkeit dramatisch mit zunehmendem Lebensalter steigt, z.B. bei Herzkreislauferkrankungen, Alzheimer, Diabetes Typ 2, Arthrose, Krebs, Sarkopenie und Osteoporose (Butler, et al. 2008; Kaeberlein 2013; Pitt and Kaeberlein 2015). Gleichzeitig sind diese Krankheiten

heutzutage auch selbst die verbreitetste Todesursache in den Industriegesellschaften (Butler, et al. 2008; Kaeberlein 2013). Analog zur Lebensspanne wird daher der Zeitraum als Gesundheitsspanne bezeichnet, in dem ein Mensch in weitgehender Abwesenheit von Schmerz sowie körperlichen und geistigen Einschränkungen fähig ist autonom zu leben (Kaeberlein 2013; Palacios, et al. 2015). Während der – meinem Eindruck nach – größere Teil der Literatur die Auffassung vertritt, die Gesundheitsspanne habe in den letzten Jahrzehnten in zumindest ähnlichem Tempo wie die Lebensspanne zugenommen (Schoeni, et al. 2008; Christensen, et al. 2009; Vaupel 2010; Fries, et al. 2011; Olshansky 2015; Palacios, et al. 2015), ist eine Minderheit der Ansicht sie hätte sich nicht wesentlich verändert (Braunseis, et al. 2012) oder sogar abgenommen (Crimmins and Beltran-Sanchez 2011; Galenkamp, et al. 2013). Aus gesellschaftlicher Sicht wäre es aus zwei Gründen wünschenswert, dass das Tempo der Erhöhung der Gesundheitsspanne mindestens mit dem der Erhöhung der Lebensspanne mithält: Erstens ist ein langes Siechtum für ein Land mit einem hochentwickelten und solidarischen Gesundheitssystem sehr kostenintensiv. Zweitens kann eine verlängerte Arbeitsfähigkeit dazu beitragen, dass die Wirtschaftskraft eines Landes trotz einer im Durchschnitt älter werdenden Bevölkerung erhalten bleibt (Sierra, et al. 2009; Vaupel 2010; Goldman, et al. 2013).

Doch sind die in den letzten Jahrzehnten beobachteten Verlängerungen der Lebens- und Gesundheitsspannen, auch ein Hinweis darauf, dass wir heute langsamer altern als frühere Generationen? Das wird nur von sehr wenigen Autoren angenommen (Vaupel and Lundstrom 1994; Vaupel 2010). Die Mehrheit führt das Phänomen vor allem auf intensive medizinische Eingriffe und Verbesserungen der Hygiene sowie des öffentlichen Gesundheitswesens zurück, die schwere Krankheitsverläufe in immer spätere Lebensjahre verschoben hätten (Wilmoth 2000; Parker and Thorslund 2007; Butler, et al. 2008; Schoeni, et al. 2008; Crimmins and Beltran-Sanchez 2011; Kaeberlein 2013; Olshansky 2015). Dass diese herkömmliche Strategie vor allem auf Fortschritte in der medizinischen Behandlung alternsbedingter Krankheiten zu setzen ausreichen wird, um den Trend höherer Lebens- und Gesundheitserwartungen in Zukunft mit ähnlichem Tempo fortzusetzen, wird in der Literatur in etwa ebenso häufig prognostiziert (Wilmoth 2000; Oeppen and Vaupel 2002; Christensen, et al. 2009; Vaupel 2010) wie das Gegenteil (Harman 2001; Kaeberlein 2013; Crimmins 2015; Dong, et al. 2016). Letztere argumentieren, dass selbst wenn es gelänge einzelne der maßgeblichen, tödlichen Krankheiten vollständig auszumerzen, der Effekt auf Gesundheits- und Lebensspanne nur marginal wäre, weil für alle anderen dieser Krankheiten immer noch ein beinahe exponentieller Anstieg im letzten Lebensdrittel der Menschen zu verzeichnen wäre. Sie plädieren daher dafür die Strategie zu ändern und gezielt den gemeinsamen Nährboden – das Altern an sich – zu bekämpfen, um so alle hauptsächlichen Todesursachen gleichzeitig hinauszuschieben (Butler, et al. 2008; Farrelly 2008; Goldman, et al. 2013; Kaeberlein 2013; Crimmins 2015). Die Verteidiger der traditionellen Methode erwidern, dass sich (i) Annahmen über maximal zu erreichende Lebenserwartungen in der Vergangenheit immer wieder als falsch erwiesen hätten (Wilmoth 2000; Oeppen and Vaupel 2002; Vaupel 2010), (ii) keine Abnahme des Trends der steigenden Lebensspannen zu beobachten sei wie es zu erwarten wäre, wenn man sich einem Maximum nähere (Wilmoth 2000; Oeppen and Vaupel 2002; Vaupel 2010) und (iii) sich Fortschritte in der Bekämpfung mehrerer Krankheiten gegenseitig synergetisch verstärken könnten (Sierra, et al. 2009; Vaupel 2010).

Der Schwerpunkt der Alternsforschung kann also eher bei der weiteren Untersuchung alternsbedingter Krankheiten oder der Erforschung des Alterns an sich gesehen werden. Während die Wirksamkeit der ersten Herangehensweise im Hinblick auf die Zukunft, wie dargelegt, umstritten ist, besteht allerdings eine bemerkenswerte Einigkeit über das Potenzial der Zweiten: Man hält es für möglich (Vaupel 2010; Crimmins 2015) bis realistisch (Harman 2001; Butler, et al. 2008; Farrelly 2008; Sierra, et al. 2009; Goldman, et al. 2013; Kaeberlein 2013; Longo, et al. 2015; Olshansky 2015), dass ein verbessertes

Verständnis des Alterns noch in den nächsten Dekaden dazu führen wird, dass man diesen Prozess wird verlangsamen können. Falls ein solches Verständnis erlangt werden kann, geht man davon aus, dass (i) sich die relative Gesundheitsspanne erhöhen bzw. das Siechtum zeitlich verdichten lassen wird (Butler, et al. 2008; Goldman, et al. 2013; Kaeberlein 2013; Longo, et al. 2015; Olshansky 2015), (ii) absolute Steigerungen der Lebens- und Gesundheisspanne erreicht werden, die ansonsten unmöglich seien (Harman 2001; Butler, et al. 2008; Goldman, et al. 2013; Kaeberlein 2013; Crimmins 2015; Olshansky 2015) und (iii) eventuell die diesbezüglichen Entwicklungen der letzten Jahrzehnte sogar in den Schatten gestellt werden (Kaeberlein 2013).

## 2. Altern und Alternsforschung

Was ist „Altern“? In der Literatur wird eine Reihe von zumeist ähnlichen Definitionen verwendet. Um an dieser Stelle eine begriffliche Basis zu schaffen, möchte ich Altern nach (Strehler 1977) wie folgt definieren: Altern umfasst sämtliche mit dem chronologischen Älterwerden verbundenen physiologischen Veränderungen eines Organismus, die (i) in allen Mitgliedern einer Population über kurz oder lang auftreten, (ii) kontinuierlich voranschreiten, (iii) nicht durch Umweltfaktoren sondern als systemimmanente Eigenschaft hervorgerufen werden und (iv) zur Abnahme der Funktionsfähigkeit und letztlich zu einer erhöhten Todeswahrscheinlichkeit beitragen. Man sollte sich dabei aber bewusst sein, dass die vom Menschen und den meisten anderen Säugetieren bekannten Charakteristiken des Alternsverlaufs – also vor allem die Abnahme der Fruchtbarkeit und Zunahme der Sterbewahrscheinlichkeit mit dem Alter – nicht universell sind und Altern im Baum des Lebens sehr unterschiedlich ausgeprägt ist (Jones, et al. 2014).

Um das Altern zu erforschen, wird versucht molekulare Marker des Alterns zu identifizieren. Ein breites Spektrum solcher Indizes, bei denen eine Korrelation mit dem chronologischen Alter beobachtet wurde, wird in der Forschung diskutiert, z.B.: sich verkürzenden Telomere, vermehrte Mutationen in der DNA, Verschleiß der Stammzellen, abnehmende Mitochondrienmasse, Zunahme ungefalteter Proteine sowie Veränderungen im Muster der Genexpression und epigenetischer Marker (Lopez-Otin, et al. 2013; Longo, et al. 2015; Pitt and Kaeberlein 2015) etc. Letztlich gibt es bislang aber keinen allgemein akzeptierten molekularen Marker des Alterns (Pitt and Kaeberlein 2015).

Die Methoden der Alternsforschung sind bei weitem zu vielfältig, um sie im Rahmen dieser Arbeit erschöpfend zu behandeln. Daher möchte ich mich darauf beschränken einige Ansätze zu nennen, die m.E. viele der verwendeten Methoden abdecken:

Es wird erforscht, ob und wie bestimmte Eingriffe die Alternsindizes ganzer biologische Systeme (Organismen, Organe, Gewebe, Zellen) verändern. Das betrifft bspw. die lebensverlängernden Effekte der Kalorienrestriktion (Carmona and Michan 2016), die Behandlung mit lebensverlängernden Substanzen wie Metformin (Onken and Driscoll 2010) oder Rapamaycin. Letzteres hat bspw. in der Maus zu durchschnittlichen Erhöhungen der Lebenspanne von 10-26% geführt (Miller, et al. 2014) und alternsbedingte Krankheiten hinausgeschoben (Johnson, et al. 2013). Weiterhin sind genetische Eingriffe zu nennen, bei denen einzelne Gene ausgeschaltet, gezielt in ihrer kodierenden Sequenz verändert oder über- bzw. unterexprimiert werden. Für hunderte Gene wurde gezeigt, dass ihre künstliche Veränderung in einer Verlängerung der Lebensspanne resultiert – allerdings zum überwiegenden Teil in Wirbellosen (Fontana, et al. 2010; Vaupel 2010).

Ein Ansatz mit ähnlicher Zielstellung ist es, mögliche Ursachen natürlicherweise auftretender Unterschiede in der Alternsrate zu identifizieren. So wird bspw. im Rahmen genomweiter Assoziationsstudien nach Kandidatengenen gesucht, in denen bestimmte Allele statistisch gehäuft bei

Individuen auftauchen, die sich als besonders lang- oder kurzlebig erwiesen haben. Bei Studien am Menschen tauchen bspw. Varianten des Gens *APOE* immer wieder als signifikanter Faktor für unterschiedliche Lebensspannen auf (Broer, et al. 2015). Weitere Arbeiten haben statt der *intra*-Spezies-Variation die *inter*-Spezies-Variation zur Grundlage, vergleichen also lang- mit kurzlebigen Arten. (Jobson, et al. 2010; Semeiks and Grishin 2012; Seim, et al. 2013; Valenzano, et al. 2015). Da dies ebenfalls Thema meiner Arbeit ist, werde ich im nächsten Kapitel noch einmal genauer darauf eingehen. Die gezielte Erforschung von Arten mit extremen Alternsraten fällt ebenfalls in diese Kategorie. Beispiele dafür sind Untersuchungen an kurzlebigen Prachtgrundkärpflingen (wenige Monate), bestimmten langlebigen Sandgräbern (30 Jahre), Fledermäusen (40 Jahre), sowie Arten, die überhaupt nicht zu altern scheinen wie Süßwasserpolypen (*Hydra*) und einigen Plattwürmern (Finch 2009; de Magalhaes 2015; Valenzano, et al. 2017). Auf Prachtgrundkärpflinge und Sandgräber wird im vierten Kapitel noch einmal eingegangen.

Eine weitere Möglichkeit besteht darin, chronologisch junge und alte Individuen derselben Art zu vergleichen, z.B. durch Analysen der Regenerationskapazität von hämatopoetischen Stammzellen (Chen, et al. 2000), des Transkriptoms (Baumgart, et al. 2014), Epigenoms (Steegenga, et al. 2014) oder Proteoms (Ori, et al. 2015) mit mehreren Alterszeitpunkten. Von den genannten Ansätzen ist dies der am stärksten deskriptive. Die meisten der in der Diskussion befindlichen Biomarker des Alterns stammen aus solchen Vergleichen.

Es ist zum gegenwärtigen Zeitpunkt noch weitgehend ungeklärt, welche exakten molekularen Ursachen dem Altern zugrunde liegen. Wie in den vorhergehenden Absätzen angedeutet, bedeutet das keineswegs, dass nicht bereits eine Vielzahl von Erkenntnissen und Ansatzpunkten für ein tieferes Verständnis des Alterns existieren würden. Nicht wenige davon werden jedoch gegenwärtig kontrovers debattiert, unterschiedlich interpretiert oder scheinen einander zu widersprechen. Der folgende Überblick konzentriert sich auf jene Punkte, die in den in meiner kumulativen Promotion zusammengefassten Manuskripten angesprochen werden:

Das Immunsystem wird mit dem Alter dysfunktional. Die Abnahme der Effizienz, was die Beseitigung von Pathogenen und beschädigten Zellen anbelangt (Licastro, et al. 2005), fällt mit der Deregulation von entzündungsbefördernden Zytokinen zusammen (Salminen, et al. 2012). Es kommt vermehrt zu chronischen Entzündungen, die – da Immunantworten sich meist nicht völlig spezifisch nur gegen Pathogene richten – zu einer Akkumulation von Gewebeschäden führen und einen Hauptrisikofaktor für alternsbedingte Krankheiten wie Alzheimer, Krebs, Arthritis und Diabetes darstellen (Chung, et al. 2009). Das Dämpfen des Entzündungssystems, z.B. durch Inhibieren des Hauptregulators NF-κB, führte in Mäusen hingegen zu einer Reduktion von Alternserscheinungen wie DNA-Schäden oder verminderter Zellproliferation (Tilstra, et al. 2012). Außerdem hat Kalorienrestriktion, als einer der prominentesten lebensverlängernden Eingriffe eine entzündungshemmende Wirkung (Carmona and Michan 2016).

Reaktive Sauerstoffspezies (englisch: *reactive oxygen species*, ROS) können in hohen Konzentrationen oxidativen Stress, d.h. Schäden an zellulären Komponenten wie DNA, Lipiden, Proteinen und insbesondere den Mitochondrien, verursachen. Sie müssen daher mit zellulären Antioxidantien wie SOD und TXN in Schach gehalten werden (Mittal, et al. 2014). Weiterhin dienen ROS in niedrigeren Konzentrationen als Signalmolekül in einer Reihe alternsrelevanter Signalwege wie der Immunantwort (West, et al. 2011), der Zellproliferation (Schieber and Chandel 2014), der Autophagie (Filomeni, et al. 2015) sowie der Apoptose (Hekimi, et al. 2016). Aufgrund dieser vielschichtigen Rollen ist es in der Literatur inzwischen umstritten, ob ROS das Altern insgesamt befördern (Edrey and Salmon 2014; Kim, et al. 2015; Davalli, et al. 2016) oder ihm entgegenwirken (Ristow and Schmeisser 2011; Schieber and

Chandel 2014; Hekimi, et al. 2016; Meng, et al. 2017). Viele Experten gehen davon aus, dass ROS sich je nach Dosis (Hormesis), Gewebe, Alter und Umwelteinflüssen unterschiedlich auf den Alternsprozess auswirken können (Labunskyy and Gladyshev 2013; Edrey and Salmon 2014; Correia-Melo and Passos 2015; Cunningham, et al. 2015; Chandrasekaran, et al. 2017).

Die Nährstoffsensor-Signalwege nehmen eine wichtige Rolle in der Alternsforschung ein, da sie entscheidend an der Vermittlung der lebensspanne- und gesundheitsfördernden Effekte der Kalorienrestriktion beteiligt sind. Wichtige Ansatzpunkte der Alternsforschung in diesem Zusammenhang sind die AMPK-, Sirtuin-, IIS- und mTOR-Signalwege. Im Zentrum des Letzteren steht ein *mechanistic target of rapamycin* (mTOR) genannter zellulärer Schlüsselregulator, der den Nährstoffpegel mit nahezu allen Aspekten des anabolischen Metabolismus verknüpft (McCormick, et al. 2011; Laplante and Sabatini 2012) (Abbildung 2). Während die mTOR-Aktivität bei Kalorienrestriktion reduziert ist, zählt seine künstliche Inhibierung mittels Rapamycin mit Nachweisen in Hefen, Fadenwürmern, Fliegen und Mäusen zu den am besten belegten lebensverlängernden Eingriffen (Johnson, et al. 2013; Carmona and Michan 2016). Die Unterexpression von mTOR durch genetische Eingriffe verlängerte die Lebensspannen der genannten Spezies (Pitt and Kaeberlein 2015) in vergleichbarer Weise – bei Mäusen bspw. um 14% (Lamming, et al. 2012). Außerdem hat die Verringerung des mTOR-Gehalts eine verzögernde Wirkung auf mehrere alternsbedingte Störungen wie etwa Herzkreislauf-, Nieren und neurodegenerative Erkrankungen wie Alzheimer als auch einige Krebsarten (Johnson, et al. 2013; Carmona and Michan 2016). Aus diesen Gründen sind sehr viele klinische Studien zu den Auswirkungen der mTOR-Inhibition auf dem Weg oder bereits abgeschlossen (Johnson, et al. 2013) (https://clinicaltrials.gov/ct2/results?term=rapamycin&Search=Search).

Die Mitochondrien stehen aus mehreren Gründen im Fokus der Alternsforschung, u.a. weil sie als Hauptproduzenten von ATP der Dreh- und Angelpunkt des zellulären Energiestoffwechsels und damit eng mit den mTOR- und IIS-Signalwegen verbunden sind (Bratic and Larsson 2013; Kotiadis, et al. 2014). Gleichzeitig sind sie die Hauptquellen von ROS, die als Nebenprodukte bei der ATP-Herstellung entstehen. Die Masse und Leistungsfähigkeit der Mitochondrien sowie die Expression von Teilen der mitochondrialen Translationsmaschinerie nimmt bei Säugetieren mit dem Alter ab, während der ROS-Ausstoß zunimmt (Bratic and Larsson 2013). Wird die mitochondriale Translationsmaschinerie zusätzlich inhibiert, verlängert das die Lebensspanne von Fadenwürmern (Houtkooper, et al. 2013), Fliegen (Copeland, et al. 2009) und Prachtgrundkärpflingen (Baumgart, et al. 2016). Die ATP-Produktion und der ROS-Ausstoß werden u.a. von den Mengenverhältnissen der Untereinheiten der Atmungskette zueinander beeinflusst (Copeland, et al. 2009; Houtkooper, et al. 2013; Miwa, et al. 2014). Da die fünf Komplexe der Atmungskette aus zahlreichen Proteinen zusammengesetzt werden, die z.T. im Kerngenom, z.T. im mitochondrialen Genom kodiert werden, hängen die Mengenverhältnisse wiederum entscheidend von der Koordinierung der mitochondrialen und kerngenomischen Genexpression ab (mitonukleäres Gleichgewicht). Deren Feinjustierung wird daher als artenübergreifender Langlebigkeitsmechanismus angesehen (Dillin, et al. 2002; Houtkooper, et al. 2013; Lionaki, et al. 2016).

Die Aktivität der beiden wichtigsten Mechanismen zur zellulären Selbstreinigung, Selbsterneuerung und Qualitätskontrolle, der lysosomalen Autophagie und des Ubiquitin-Proteasom-Systems, nimmt mit dem Alter ab (Saez and Vilchez 2014; Martinez-Lopez, et al. 2015). Gleichzeitig gilt die Zunahme beschädigter Organellen, insbesondere der Mitochondrien, sowie abnormaler Proteine als eines der Merkmale des Alterns (Lopez-Otin, et al. 2013; Carmona and Michan 2016). Die Akkumulation beschädigter oder falsch- bzw. ungefalteter Proteine ist außerdem mit alternsbedingten neurodegenerativen Erkrankungen assoziiert, bspw. Amyloid-Beta mit Alzheimer und α-Synuclein mit

Parkinson (Rubinsztein, et al. 2011; Carmona and Michan 2016). In ähnlicher Weise spielen nicht-enzymatisch, durch Kohlenhydrate modifizierter Proteine (*advanced glycosylation end products*, AGEs) eine Rolle bei der Pathogenese verschiedener Krankheiten, z.B. bei Diabetes und chronischem Nierenversagen (Vistoli, et al. 2013).

Die mRNA-Translation ist ein energieaufwändiger Prozess, der ca. 30-40% des gesamten ATP-Haushalts verschlingt und der notwendig für Wachstum als auch Proliferation ist (Hands, et al. 2009). Eine Reihe von Eingriffen, die u.a. die Translation global hemmen, bspw. Kalorienrestriktion und mTOR-Inhibition, verlängern die Lebensspanne von Hefen, Fliegen, Fadenwürmern und Mäusen. In Wirbellosen wurde gezeigt, dass die Unterexpression verschiedener Translationsinitiationsfaktoren und ribosomaler Proteine die Lebensspanne verlängert (Johnson, et al. 2013). Passend dazu ist die Expression ribosomaler Proteine in Prachtgrundkärpflingen negativ mit der Lebensspanne korreliert (Baumgart, et al. 2016). Der Transkriptionsfaktor MYC reguliert 15-20% aller Gene in Säugetieren, wobei der größte Teil der hochregulierten Gene in die Proteintranslation involviert ist. Unterexpression von MYC in der Maus, führte u.a. zu einer signifikant niedrigen Translationsrate und verlängerte die Lebensspanne von Mäusen um 15% (Hofmann, et al. 2015). Insgesamt deutet also vieles daraufhin, dass ein globales Absenken der Translationsrate sich positiv auf die Lebensspanne auswirkt.

Aus diesen und vielen weiteren Erkenntnissen, aber auch bloßen Vermutungen wurden zahlreiche Alternstheorien konstruiert. So wurden schon 1990 mehr als 300 Theorien gezählt, die im Rahmen wissenschaftlicher Publikationen formuliert worden waren (Medvedev 1990). Man kann dabei zwischen mechanistischen und evolutionären Alternstheorien unterscheiden. Mechanistische Alternstheorien versuchen die Gründe für das Altern auf molekularer, zellulärer und Organ-Ebene zu erklären. So postuliert etwa die Hyperfunktionstheorie des Alterns – ausgehend von den oben beschriebenen Erkenntnissen um den mTOR-Signalweg – dass das in der Jugend noch nützliche Wachstumsprogramm im Erwachsenenalter nicht vollständig abgeschaltet werde. Das daraus folgende übermäßige Zellwachstum sowie die zelluläre Hyperaktivität würden mit der Zeit die innere Ordnung der Gewebe und Organe untergraben und so zu den typischen alternsbedingten Krankheiten führen (Blagosklonny 2012; Gems and Partridge 2013) (Abbildung 2).

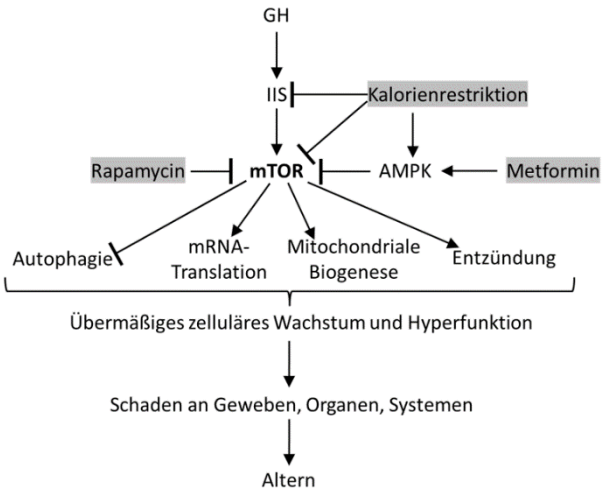



**Abbildung 2.** Hyperfunktionstheorie des Alterns. Die Aktivierung des mTOR-Signalwegs nach Abschluss der körperlichen Entwicklung entspricht der Theorie nach einem „pervertierten" Wachstum, das auf der Makroebene nach und nach zu einem Verlust der Homöostase führe. Die Eingriffe, die mTOR inhibieren und zu einer Verlängerung der Lebensspanne führen (grau unterlegt), unterdrücken nach dieser Lesart die schädliche Fortsetzung des Wachstumsprogramms. GH – *Growth hormone* (Wachstumshormon), IIS – *Insulin/insulin-like growth factor-1* Signalweg, AMPK –AMP-aktivierte Proteinkinase.

Evolutionäre Alternstheorien gehen demgegenüber der Frage nach dem ultimativen „warum“ nach. Warum sind wir nicht unsterblich? Obwohl die Antworten der Evolutionsbiologen im Detail z.T. auseinandergehen, sind sie sich im Hinblick auf den letztendlichen Grund für das Altern doch weitgehend einig: Lebewesen verfallen deshalb aus sich selbst heraus und sterben schließlich daran, weil sie auch aus anderen Gründen außer dem eben genannten sterben – Fressfeinde, nicht alternsbedingte Krankheiten, Erfrieren, Verhungern etc. Bspw. können Mäuse bis zu 4 Jahre alt werden (Tacutu, et al. 2013). In der freien Natur hingegen erreichen nicht mehr als 10% der Mäuse das zweite Lebensjahr, wobei von den restlichen 90% nur wenige am Altern sterben dürften (Phelan and Austad 1989). Da unabhängig vom Altern nur wenige Individuen ein hohes Alter erreichen, nimmt die Kraft der natürlichen Selektion in Bezug auf spätere Lebensphasen immer weiter ab (Williams 1957; Gems and Partridge 2013). Die Theorie der antagonistischen Pleiotropie geht von genetischen Merkmalen aus, die in der Jugend vorteilhaft sind und im Alter schädliche Effekte haben. Wenn die Wirkung in der Jugend für die evolutionäre Fitness überwiegt, was nach den oben beschriebenen Überlegungen zur extrinsischen Mortalität häufig der Fall ist, würden diese Mutationen im Ergebnis positiv selektiert, d.h. sich in der Population durchsetzen (Williams 1957). Die oben erwähnte Hyperfunktionstheorie ergänzt die evolutionäre Alternstheorie der antagonistischen Pleiotropie, um einen konkreten Mechanismus, indem sie sie mit den empirischen Befunden um den mTOR-Signalweg verbindet. Zudem wird die antagonistische Pleiotropie dadurch gestützt, dass ihre Vorhersage eines Zielkonflikts zwischen Fitness in frühen und späten Lebensphasen in vielerlei Hinsicht belegt wurde. So zeigten Selektionsexperimente an Fruchtfliegen (Hughes and Reynolds 2005), Assoziationsstudien an vielen Spezies, u.a. dem Menschen (Pitt and Kaeberlein 2015), und genetische Eingriffe an Wachstumsfaktoren von Mäusen (Bartke 2012), regelmäßig eine negative Korrelation zwischen Wachstumsrate und früher Fruchtbarkeit auf der einen Seite sowie der Lebensspanne auf der anderen Seite.

## 3. Positive Selektion und der Zweig-Positionstest

Meine Arbeit bedient sich vorwiegend einer Methode aus dem Feld der komparativen Genomik (bzw. Transkriptomik) – basiert also auf dem Vergleich von Genom- bzw. Transkriptomsequenzen verschiedener Organismen. Grundsätzlich wird in diesem Feld das Muster von evolutionär konservierten, d.h. ähnlichen oder unveränderten Bereichen, auf der einen und Unterschieden auf der anderen Seite untersucht. Die Entwicklung der komparativen Genomik ist eng mit den Fortschritten der Sequenziertechnologie verwoben, die die Kosten und die Zeit, die benötigt werden, um Genome bzw. Transkriptome zu sequenzieren, seit dem Humangenomprojekt (Lander, et al. 2001) drastisch reduziert haben. Damit sind nicht nur die Möglichkeiten einzelner Arbeitsgruppen enorm gestiegen, Organismen zu sequenzieren, um basierend darauf Vergleiche zu tätigen, sondern auch die Menge öffentlich verfügbarer Daten rasant gewachsen (Alfoldi and Lindblad-Toh 2013). Von beidem hat meine Arbeit stark profitiert. Mit dem Zuwachs an Daten ging eine dynamische Neu- und Weiterentwicklung der Methoden einher, die zu einer großen Vielfalt der Anwendungsgebiete der komparativen Genomik geführt hat. Sie wird heute u.a. eingesetzt um Homologiebeziehungen zwischen Genen zu ermitteln (Li, et al. 2003; Hou, et al. 2016; Petersen, et al. 2017), funktionell relevante Bereiche wie etwa Transkriptionsfaktorbindestellen vorherzusagen (Chen, et al. 2004; Defrance and Touzet 2006; Williams, et al. 2016), phylogenetische Speziesbäume zu rekonstruieren (Blanga-Kanfi, et al. 2009; Rowe, et al. 2010; Ji, et al. 2017) u.v.m. Insbesondere geht es bei der komparativen Genomik häufig darum, jene genomischen Bereiche und Varianten zu identifizieren, die phänotypischen Unterschieden bzw. Gemeinsamkeiten zwischen verschiedenen Spezies oder zwischen Individuen derselben Spezies zu Grunde liegen könnten.

Eine evolutionsbiologische Methode in diesem Zusammenhang ist die Suche nach positiv selektierten Genen (PSGs), die die wesentliche Grundlage für meine Arbeit darstellt. An dieser Stelle sollen zunächst die Begriffe positive bzw. negative Selektion definiert werden:

- Positive Selektion bezeichnet das Phänomen, dass sich ein neues Allel über die Generationen hinweg in einer Population bzw. einer Spezies verbreitet, weil es die evolutionäre Fitness erhöht – d.h. die Fähigkeit zu überleben und sich zu reproduzieren stärkt. Am Ende eines solchen Ausbreitungsprozesses hat das alternative Allel das ursprüngliche Allel der Population/Spezies am entsprechenden Lokus verdrängt. Da dieses Konzept auf genetischer Ebene widerspiegelt, was in Darwins epochalem Werk „Über die Entstehung der Arten“ als Kernmechanismus der langsamen Anpassung der Spezies an ihre Umwelt beschrieben wird (Darwin 1859), wird in der Literatur z.T. auch von gerichteter Darwin’scher Selektion (Wang, et al. 2006; McClellan 2013; Sheng, et al. 2016) bzw. adaptiver Selektion (Wu, et al. 2014) gesprochen.
- Negative oder reinigende Selektion beschreibt den Vorgang des Entfernens von für die evolutionäre Fitness nachteiligen/schädlichen Allelen aus einer Population, weil die Träger dieser Allele weniger Nachkommen haben als andere Individuen.
- Mutationen, die keinen Effekt auf die Fitness haben, sind keinem Selektionsdruck ausgesetzt; man sagt auch sie stehen unter neutraler Selektion. Die Frequenz entsprechend mutierter Allele in der Population kann sich durch Zufall ändern (genetische Drift).

Es existieren verschiedene Ansätze zur Identifizierung positiver Selektion, die hinsichtlich der untersuchten evolutionären Distanzen bzw. der jeweils dazu verwendeten Daten klassifiziert werden können. Methoden wie Tajima’s D (Tajima 1989), Testen auf Kopplungsungleichgewichte (Sabeti, et al. 2002) oder $F_{ST}$ (Lewontin and Krakauer 1973) fahnden auf Basis von Intra-Spezies Polymorphismen nach – aus evolutionärer Sicht – kurz zurückliegenden Ereignissen positiver Selektion. Der Hudson-Kreitman-Aguadé-Test (Hudson, et al. 1987) und der McDonald-Kreitman-Test (McDonald and Kreitman 1991) schließen aus der Variation innerhalb einer Spezies im Vergleich mit der Variation zwischen den Spezies auf positive Selektion. Bei reinen Speziesvergleichen – wie in meiner Arbeit – kommen i.d.R. Methoden zum Einsatz, die auf dem $d_N/d_S$-Verhältnis (auch $K_a/K_s$ oder $\omega$) beruhen und mit denen auch lange zurückliegende Ereignisse positiver Selektion detektiert werden können, was ansonsten nur durch das Sequenzieren alter DNA aus Fossilen möglich wäre (Biswas and Akey 2006; Fu and Akey 2013; Wu, et al. 2014).

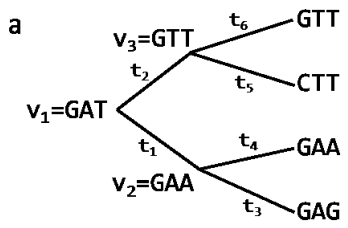
$d_N$ steht dabei für die Rate nicht-synonymer (d.h. Aminosäure-ändernder) und $d_S$ für die Rate synonymer (d.h. Aminosäure- nicht-ändernder) Substitutionen. Das $d_N/d_S$-Verhältnis kann also nur in proteinkodierenden Sequenzen berechnet werden und daher sind darauf aufbauende Methoden auf diese Genombereiche beschränkt. Promotoren und andere für die Genexpression relevante Sequenzabschnitte können bspw. mit diesen Methoden nicht untersucht werden. Dass Codons mutiert werden können, ohne dass sich dadurch die kodierte Aminosäure ändert, folgt aus der Degeneriertheit des genetischen Codes – 64 Codons kodieren zusammen für nur 20 kanonische Aminosäuren. Die beobachteten nicht-synonymen Substitutionen zwischen den Spezies können sowohl das Produkt positiver Selektion als auch ein Produkt genetischer Drift und mithin neutral selektiert sein – das ist das Grundproblem bei der Suche nach positiver Selektion. Von den synonymen Substitutionen in kodierenden Sequenzen wird angenommen, dass sie ein Maß für neutrale Selektion darstellen, bezüglich dessen entschieden werden kann, ob $d_N$ beschleunigt oder entschleunigt ist bzw. ob sich nicht-synonyme Substitutionen mit höherer oder niedriger Wahrscheinlichkeit durchsetzen als solche unter neutraler Selektion (Miyata and Yasunaga 1980; Nielsen 2001; Yang 2005; Biswas and Akey 2006; Fu and Akey 2013):

- $d_N > d_S$ → positive Selektion
- $d_N < d_S$ → negative Selektion
- $d_N = d_S$ → neutrale Selektion

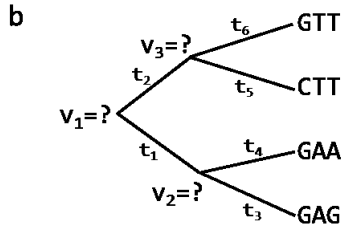
$d_N$ und $d_S$ werden auf der Grundlage zweier gleich langer proteinkodierenden Sequenzen (englisch: coding sequence, CDS) berechnet, indem man die Zahl der beobachteten nicht-synonymen bzw. synonymen Austausche zwischen diesen Sequenzen ins Verhältnis setzt zu den möglichen nicht-synonymen bzw. synonymen Substitutionen und abschließend für Rückmutationen korrigiert (Miyata and Yasunaga 1980; Nei and Gojobori 1986). Dieser einfache Ansatz bietet allerdings so gut wie keine Sensitivität, weil die meisten Proteine große Regionen enthalten, in denen fast keine Austausche toleriert werden und $d_N$ nahe 0 ist, während i.d.R. nur wenige Positionen jemals unter positiver Selektion stehen, sodass diese Signale beim Mitteln über die Positionen einer CDS in der Masse negativer Selektion untergehen (Nielsen 2001; Kosakovsky Pond and Frost 2005; Yang 2005). Ein weiteres Problem ist, dass diese Methode ungeeignet ist, um positiv selektierte Gene (PSGs) in einzelnen Spezies bzw. auf Zweigen einer Phylogenie zu identifizieren, weil sie dazu die explizite Rekonstruktion anzestraler Sequenzen benötigt, die aber mit zufälligen Fehlern und systematischen Verzerrungen behaftet ist (Collins, et al. 1994; Yang, et al. 1995).

Um diese Probleme zu umgehen, wurden auf Grundlage des $d_N/d_S$-Verhältnisses aufwändigere Tests entwickelt, die in Programmpakete wie PAML (Yang 1997, 2007) und HyPhy (Pond, et al. 2005) implementiert wurden. Für diese Tests wird eine Alignierung von i.d.R. mindestens drei Sequenzen und ein dazugehöriger phylogenetischer Baum benötigt. Der Zweig-Positionstest (englisch *branch-site test*) ermöglicht es positive Selektion gezielt auf einzelnen Zweigen dieses Baums zu detektieren und auch dann, wenn sie nur auf einzelnen Positionen einer CDS wirkt (Yang and Nielsen 2002; Zhang, et al. 2005).

Beim Zweig-Positionstest wird die Wahrscheinlichkeit dafür berechnet, dass die gegebene Alignierung als Ergebnis eines Evolutionsverlaufs auf dem ebenfalls gegebenen Baum zustande kommt. Die zentrale Idee dabei ist, das zwei Mal auf Basis von Substitutionsmodellen durchzuführen, die sich in genau einem Punkt unterscheiden: Das eine Mal wird positive Selektion ($d_N > d_S$) auf dem zur Untersuchung ausgewählten Zweig der Phylogenie als Möglichkeit zugelassen (Alternativszenario $H_A$), das andere Mal nicht (Nullszenario $H_0$). Anschließend wird mit einem Wahrscheinlichkeits-Quotienten-Test überprüft, ob die errechnete Wahrscheinlichkeit für das Alternativszenario $W(H_A)$ signifikant höher ist als die des Nullszenarios $W(H_0)$ (Yang 2007). Dabei wird angenommen, dass alle Positionen der Alignierung unabhängig voneinander evolvieren. Die jeweilige Gesamtwahrscheinlichkeit $W(H_A)$ bzw. $W(H_0)$ wird dementsprechend als Produkt der Einzelwahrscheinlichkeiten aller Positionen, d.h. Spalten der Codon-Alignierung (Yang and Nielsen 2002) berechnet. Die Wahrscheinlichkeit für das Zustandekommen einer einzelnen Alignierungsspalte wird auf der Grundlage der Modellierung von Substitutionen als zeitkontinuierliche Markov-Prozesse bestimmt. Konkret bedeutet das, dass für jeden Ast mit der Astlänge $t_x$ des gegebenen Baums eine Übergangsmatrix $P(t_x)$ bestimmt wird, die für alle 61 aminosäurekodierenden Codons $i$ und $j$ – mit dem Eintrag $P_{ij}(t_x)$ – die Wahrscheinlichkeit angibt, dass $i$ auf der Länge von $t_x$ in $j$ umgewandelt wird. Wenn für jeden Knoten des Baums, d.h. für jede der heute existierenden Spezies und alle ihrer gemeinsamen Vorfahren, bekannt wäre, welches Codon der Zustand an dem entsprechenden Knoten ist bzw. war, könnte für jeden Ast mit der Länge $t_x$ die entsprechende bedingte Wahrscheinlichkeit $P_{ij}(t_x)$ bestimmt werden. Das Produkt dieser bedingten Wahrscheinlichkeiten über alle Äste entspricht dann der Wahrscheinlichkeit für den – durch den Baum und die Codons an den Knoten gegebenen – Evolutionsverlauf (siehe Abbildung 3a für ein Beispiel).




$$W = \pi_{GAT} P_{GAT,GAA}(t_1) P_{GAT,GTT}(t_2) P_{GAA,GAG}(t_3) \\ P_{GAA,GAA}(t_4) P_{GTT,CTT}(t_5) P_{GTT,GTT}(t_6)$$




$$W = \sum_{v_1 \in A} \sum_{v_2 \in A} \sum_{v_3 \in A} \pi_{v_1} P_{v_1,v_2}(t_1) P_{v_1,v_3}(t_2) P_{v_2,GAG}(t_3) \\ P_{v_2,GAA}(t_4) P_{v_3,CTT}(t_5) P_{v_3,GTT}(t_6)$$

**Abbildung 3.** Beispiele für die Berechnung der Wahrscheinlichkeit einer Codon- Alignierungsspalte nach einem Markov-Prozess-Substitutionsmodell. (a) Ein Evolutionsverlauf sei durch einen phylogenetischen Baum mit bekannten Astlängen $t_1$-$t_6$ sowie den Codons an allen seinen Knoten gegeben (links). Die Wahrscheinlichkeit $W$ dieses Evolutionsverlaufs ist nach dem Modell durch die Gleichung rechts gegeben. Die Berechnung erfolgt auf der Basis einer Übergangsmatrix $P(t_x)$, die für alle Codons $i$ und $j$ – mit dem Eintrag $P_{ij}(t_x)$ – die Wahrscheinlichkeit angebe, dass $i$ auf der Länge von $t_x$ in $j$ umgewandelt wird. $\pi_{v_1}$ gebe die Initialwahrscheinlichkeit für das Codon $v_1$ an der Wurzel an. $\pi$ kann im Rahmen des Zweigpositionstests empirisch aus den Nukleotidfrequenzen an den drei Codon-Positionen der gegebenen Sequenzen der heute existierenden Spezies bestimmt werden. (b) In der Praxis sind die Zustände an den inneren Knoten, hier $v_1$-$v_3$, i.d.R. unbekannt. Die Wahrscheinlichkeit des Evolutionsverlaufs ergibt sich dann als Summe über alle möglichen Zustände an den inneren Knoten (Gleichung rechts, $A$ sei dementsprechend die Menge der aminosäurekodierenden Codons). Die Zahl der nötigen Rechenschritte, die so wie oben dargestellt exponentiell mit der Zahl terminaler Knoten bzw. Spezies wachsen würde, kann durch Umformung der Gleichung auf einen linearen Zusammenhang reduziert werden (Felsenstein 1981).

In der Praxis ist allerdings zumeist nicht bekannt, welche Codons in den gemeinsamen Vorfahren, also an den inneren Knoten des Baums, existierten. Um die erwähnten systematischen Fehler zu vermeiden, die bei Rekonstruktion nur des jeweils wahrscheinlichsten anzestralen Zustands entstehen (Collins, et al. 1994), werden beim Zweig-Positionstest alle möglichen Zustände (d.h. alle aminosäurekodierenden Codons) an den inneren Knoten betrachtet und die Wahrscheinlichkeiten für sämtliche sich daraus ergebenden Evolutionsverläufe aufsummiert (Abbildung 3b). Die Übergangsmatrix $P(t)$ wird aus einer Ratenmatrix $Q$ bestimmt ($P(t) = e^{tQ}$), die ihrerseits durch mehrere Parameter definiert wird, von denen der wichtigste $\omega = d_N/d_S$ ist. Wie bereits erwähnt, war die Annahme, dass der Selektionsdruck und mithin $\omega$ über die Positionen der Alignierung und die Zweige der Phylogenie variiert, der wesentliche Grund für die Entwicklung des Zweig-Positionstests. Dies wird durch das Modell zum einen durch die Unterscheidung in den zur Untersuchung auf positive Selektion ausgewählten Evolutionszweig (Vordergrundzweig, VZ) und alle anderen Zweige (Hintergrundzweige, HZ) abgebildet; zum anderen durch die Annahme der Existenz von vier Positionsklassen. Die Positionsklassen beschreiben Positionen die auf beiden Zweigklassen gleichermaßen (i) negativ ($0 < \omega_1 < 1$) bzw. (ii) neutral ($\omega_2 = 1$) evolvieren sowie Positionen, die auf dem VZ unter positivem ($\omega_3 > 1$) und auf den HZ unter (iii) negativem bzw. (iv) neutralem Selektionsdruck stehen. Das Szenario $H_0$ unterscheidet sich von $H_A$ ausschließlich dadurch, dass der VZ in den letzten beiden Positionsklassen auf neutrale statt auf positive Selektion festgelegt ist. Da nicht bekannt ist zu welcher Positionsklasse eine Alignierungsspalte gehört, wird die Wahrscheinlichkeit ihres Zustandekommens unter jeder der vier Positionsklassen separat ermittelt und anschließend gemittelt – wobei die angenommenen Anteile der Positionsklassen an der Alignierung als Gewichte dienen. Über die Anteile der Positionsklassen, die konkrete Stärke des Selektionsdrucks in einigen Positionsklassen (d.h. $\omega_1$ und $\omega_3$) sowie weitere Parameter wird maximiert, d.h. es wird mit dem Ziel die Gesamtwahrscheinlichkeit $W(H_A)$ bzw. $W(H_0)$ möglichst groß zu machen

über den Parameterraum iteriert. Nach Schätzung der Parameter wird für jede Alignierungsspalte ihre Wahrscheinlichkeit bestimmt unter positiver Selektion zu stehen. Dies geschieht mit einem *Bayes empirical Bayes* genannten Verfahren auf Grundlage der Wahrscheinlichkeiten des Zustandekommens der jeweiligen Alignierungsspalte unter den vier Positionsklassen im Alternativszenario $H_A$. (vgl. (Felsenstein 1981; Goldman and Yang 1994; Yang 1998; Yang and Nielsen 2002; Yang, et al. 2005; Zhang, et al. 2005))

Simulationen haben gezeigt, dass der Zweig-Positionstest robust gegenüber verschiedenen Verletzungen der Annahmen ist, auf denen er beruht – bspw. wenn entgegen den Annahmen ebenfalls auf den HZ positive Selektion stattfindet (Zhang, et al. 2005; Gharib and Robinson-Rechavi 2013). Der Test ist allerdings nicht robust gegenüber Alignierungsfehlern, die mit hoher Wahrscheinlichkeit in falsch-positiven Ergebnissen resultieren (Fletcher and Yang 2010). Alignierungsprobleme treten bspw. gehäuft in der Nähe von Alignierungslücken und Bereichen mit hoher Sequenzdivergenz auf (Mallick, et al. 2009). Diese Probleme werden begünstigt durch Fehler, die vorab in der Prozessierung der Sequenzen, z.B. bei der Genannotation oder Orthologzuweisung, geschehen – etwa wenn Exons in einzelnen Sequenzen falsch vorhergesagt wurden. In solchen und anderen Fällen kann es dazu kommen, dass nicht-orthologe Codons in eine Alignierungsspalte gefasst werden, was i.d.R. zu einer artifiziell erhöhten Messung der Rate nicht-synonymer Substitutionen führt (Yang and dos Reis 2011). Alignierungsprobleme werden so als Signal positiver Selektion fehlinterpretiert. Aufgrund der tausenden Gene, die in einer genomweiten Analyse betrachtet werden, auf der einen Seite und der verhältnismäßig niedrigen Zahl von zu erwartenden Genen mit echten Signalen positiver Selektion auf der anderen, kann schon eine winzige Fehlerrate dazu führen, dass die identifizierten PSGs überwiegend aus Falsch-Positiven bestehen (Mallick, et al. 2009; Schneider, et al. 2009). Um das zu vermeiden, werden in nahezu allen genomweiten Suchen nach positiver Selektion Filterstrategien eingesetzt, die z.B. versuchen problematische Alignierungsregionen vor Anwendung des Tests zu entfernen Diese Filterstrategien werden wie weitere Schritte, die zur genomweiten Anwendung des Zweig-Positionstests notwendig sind, von jeder Arbeitsgruppe selbst implementiert und weisen daher eine große Spannbreite auf, was eingesetzte Methoden und Qualitätsstandards betrifft. Als Resultat dessen wurden in einigen Arbeiten hohe Falsch-Positiv-Raten nachgewiesen. So wurde etwa bei der Neuuntersuchung einer Stichprobe vermeintlich positiv selektierter Schimpansengene gezeigt, dass die Mehrzahl davon auf Sequenz- und Alignierungsartefakte zurückzuführen ist (Mallick, et al. 2009) und die Falsch-Positiv-Rate in einer Untersuchung von *Drosophila*-Fliegen auf mindestens 45% geschätzt (Markova-Raina and Petrov 2011).

Der Zweig-Positionstest wurde in zahlreichen genomweiten Analysen eingesetzt – zumeist mit dem Ziel Phänotypen, die auf bestimmten evolutionären Zweigen entwickelt wurden, mit Genen in Verbindung zu bringen, die auf diesen Zweigen positiv selektiert wurden. So wurde etwa die Fähigkeit der tibetischen Antilope in großen Höhen mit starker UV-Strahlung und niedrigen Sauerstoffkonzentrationen zu überleben mit Anreicherungen positiv selektierter Gene (PSGs) in den funktionellen Kategorien DNA-Reparatur und Energiemetabolismus assoziiert. Ebenfalls mit positiver Selektion im Bereich des Energiemetabolismus wurde die Evolution der Flugfähigkeit im Vorfahren der Fledermäuse erklärt – eine der energieaufwändigsten Tätigkeiten überhaupt (Shen, et al. 2010). Weiterhin wurden bspw. zwischen bakteriellen Krankheitserregern auf der einen sowie immunrelevanten Säugetiergenen auf der anderen Seite durch positive Selektion angetriebene „Rüstungswettläufe“ dokumentiert (Petersen, et al. 2007; Webb, et al. 2015). Der Mensch wurde von einer ganzen Reihe von Arbeiten genomweit auf positive Selektion hin untersucht (z.B. (Bakewell, et al. 2007; Kosiol, et al. 2008; Gaya-Vidal and Alba 2014)). Mit Blick auf das Thema dieser Arbeit ist zu sagen, dass Studien, die die Genomsequenzen besonders langlebiger Spezies veröffentlicht haben, regelmäßig auch eine Suche nach PSGs auf Basis des Zweig-

Positionstests beinhalteten. So etwa beim Blindmull (>20 Jahre, (Fang, Nevo, et al. 2014)), beim Nacktmull (> 30 Jahre, (Kim, et al. 2011)), der großen Bartfledermaus (> 40 Jahre, (Seim, et al. 2013)) und beim Grönlandwal (> 200 Jahre, (Keane, et al. 2015)). Dabei wurden u.a. PSGs identifiziert, die für die Verlängerung von Telomeren entscheidend (Kim, et al. 2011), mit Entzündungen bzw. dem Immunsystem (Seim, et al. 2013; Fang, Nevo, et al. 2014) oder mit Krebs und anderen Krankheiten assoziiert sind (Fang, Nevo, et al. 2014; Keane, et al. 2015).

## 4. Untersuchte Spezies und die Evolution ihrer Lebensspannen

Im vorangegangenen Kapitel wurde dargelegt, dass durch die Suche nach positiver Selektion Gene identifiziert werden können, durch deren Veränderung im Laufe der Evolution eine Veränderung phänotypischer Eigenschaften einer Spezies vermittelt wurde. Meine Arbeit hatte das Ziel Gene und Positionen in Genen zu finden, die relevant für das Altern der entsprechenden Spezies sein könnten. Daher fahndete ich, ähnlich den Studien die zum Ende des vorhergehenden Kapitels erwähnt wurden, auf evolutionären Zweigen lang- und kurzlebiger Spezies nach positiver Selektion, auf denen sich die Lebenspanne aller Wahrscheinlichkeit nach verändert hat. Dazu bot es sich an, den Fokus auf Spezies zu richten, die eine herausragend kurze oder herausragend lange Lebensspanne haben und deren Sequenzen mit denen von Spezies zu vergleichen, die weniger extreme Werte aufweisen. Konkret ging es in meiner Arbeit um die kurzlebige Gattung der Prachtgrundkärpflinge und die langlebigen Vertreter der Sandgräber.

Die Gattung der Prachtgrundkärpflinge (*Nothobranchius*) besteht aus Fischarten, die wenige Zentimeter groß werden und in Äquatorial- und Subäquatorialafrika beheimatet sind (Abbildung 4a/b). Dort leben sie in temporär existierenden Teichen oder Tümpeln, die meist nur während der Regenzeit Wasser führen. Während die erwachsene Population in der Trockenzeit i.d.R. ausnahmslos zu Grunde geht, sind die Eier in der Lage diese in einem Zustand angehaltener Entwicklung zu überstehen. Die nächste Generation schlüpft zu Beginn der folgenden Regenzeit und setzt den Zyklus fort (Jubb 1981; Cellerino, et al. 2016). Werden die Tiere im Aquarium gehalten, bleibt die Lebensspanne einiger Arten trotz des Wegfalls der hohen, klimabedingten extrinsischen Mortalität limitiert. So erreicht der in dieser Hinsicht extremste Vertreter der Gattung, der Türkise Prachtgrundkärpfling (*Nothobranchius furzeri*), auch in Gefangenschaft nur eine durchschnittliche Lebensspanne von 6 Monaten, wobei er dort einen typischen Alternsprozess zeigt (Di Cicco, et al. 2011) und einige seiner Stämme sogar noch niedrigere Lebenserwartungen aufweisen. Damit gilt der Türkise Prachtgrundkärpfling als kurzlebigste Wirbeltierspezies überhaupt, die in Gefangenschaft gehalten werden kann (Valdesalici and Cellerino 2003; Lucas-Sanchez, et al. 2014). Der Türkise Prachtgrundkärpfling stammt wie andere *Nothobranchius*-Arten mit besonders schnellem Wachstum und besonders kurzer Lebensspanne aus einer sehr trockenen Region. Demgegenüber sind etwas langlebigere Arten, die in Aquariumshaltung im Durchschnitt bis zu 18 Monate alt werden können, in Gebieten mit längeren Regenzeiten beheimatet (Tozzini, et al. 2013). Es gilt als wahrscheinlich, dass umwälzende, historische Klimaveränderungen, die die verschiedenen Lebensräume der Prachtgrundkärpflinge in unterschiedlicher Weise betrafen, die Diversifikation der Gattung entscheidend vorangetrieben haben. Insbesondere geht man davon aus, dass eine langanhaltende Periode zunehmender Trockenheit in Ostafrika die jährlich wiederkehrende hohe extrinsische Mortalität verursacht und damit einen starken Selektionsdruck auf schnelles Wachstum und frühe Fruchtbarkeit erzeugt hat. Das beschleunigte Altern, das für die evolutionäre Fitness aus dem gleichen Grund – der hohen extrinsischen Mortalität – weitgehend bedeutungslos sein dürfte, wird als Kehrseite angesehen (Valdesalici and Cellerino 2003; Dorn, et al. 2014; Cellerino, et al. 2016) (siehe auch evolutionäre Alternstheorien am Ende des zweiten Kapitels). Diese Periode der Trockenheit begann vor etwa 8 Mio.

Jahren und hält mit mehreren kurzen Unterbrechungen feuchteren Klimas bis heute an (Trauth, et al. 2005; Sepulchre, et al. 2006). Vor ebenfalls ca. 8 Millionen Jahren entstand aus einer sehr wahrscheinlich noch nicht einjährigen und daher deutlich langlebigeren Spezies der letzte gemeinsame Vorfahr der Prachtgrundkärpflinge (Dorn, et al. 2014; Furness, et al. 2015). Von den Manuskripten abgesehen, die Teil dieser Arbeit sind, hat sich in der Literatur bisher nur eine weitere Arbeit mit der genomweiten Identifizierung positiver Selektion bei Prachtgrundkärpflingen auseinandergesetzt (Valenzano, et al. 2015).

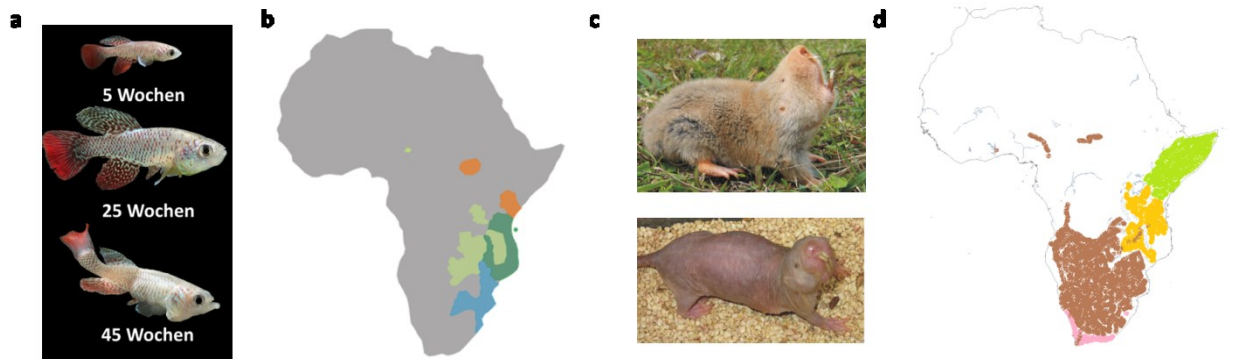



**Abbildung 4.** Untersuchte Spezies mit Verbreitungsgebieten. a) Männliche Türkise Prachtgrundkärpflinge (*Nothobranchius furzeri*) in verschiedener Alternsstufen ((Platzer, et al. 2011), bearbeitet). b) Verbreitungsgebiet der Gattung der Prachgrundkärpflinge. Die Farben zeigen die Regionen an, in denen die Vertreter der evolutionären Kladen der Gattung beheimatet sind. Der Türkise Prachtgrundkärpfling und andere kurzlebige Arten der südlichen Klade sind in der blau markierten Region beheimatet (Dorn, 2014 #498}, bearbeitet). c) Ausgewählte Spezies der Familie der Sandgräber (Bathyergidae); oben: Silbermull (*Heliophobius argenteocinerus*), unten: Nacktmull (*Heterocephalus glaber*) ((Seney, et al. 2009), bearbeitet). d) Verbreitungsgebiet der Sandgräber. Die Farben zeigen die Regionen an, in denen die Gattungen der Familie beheimatet sind: Grün – Nacktmulle (*Heterocephalus*), orange – Erdbohrer (*Heliophobius*), braun – Graumulle (*Cryptomys* und *Fukomys*), rosa: Strandgräber (*Bathyergus*) und Blessmulle (*Georychus*) (nach (Burda 2001)).

Die Sandgräber (Bathyergidae) sind eine in unterirdischen Tunnelsystemen lebende Familie der Nagetiere, die hauptsächlich im südlichen und östlichen Afrika beheimatet ist ((Bennett and Faulkes 2000), Abbildung 4c/d). Es ist wohlbekannt, dass bei Säugetieren größere und schwerere Spezies dazu tendieren länger zu leben als kleinere (de Magalhaes, et al. 2007; Fushan, et al. 2015). Die meisten Sandgräberarten liegen mit einem Körpergewicht von durchschnittlich 35-250 g im Größenfeld etwa zwischen Mäusen und Ratten. Während Letztere allerdings höchstens vier Jahre alt werden, haben sämtliche bislang untersuchten Sandgräberarten eine maximale Lebensspanne von mehr als zehn Jahren (Tacutu, et al. 2013). Der Nacktmull, als extremster Vertreter der Familie, kann sogar älter als 30 Jahre werden, zeigt darüber hinaus kaum Alternserscheinungen und besitzt eine bemerkenswerte Krebsresistenz (Lagunas-Rangel and Chavez-Valencia 2017). Der Nacktmull und einige andere Sandgräber haben eine deutlich höhere Lebenserwartung als aufgrund ihres Gewichts zu erwarten wäre (Buffenstein 2008; Tacutu, et al. 2013). Demgegenüber gilt für den nächsten Verwandten der Familie – das ist die Rohrratte – mit einer maximalen Lebensspanne von ca. 5 Jahren und einem mittleren Gewicht von 7 kg das Gegenteil (Tacutu, et al. 2013) (Abbildung 5). In der Literatur werden zwei mögliche Gründe angeführt, die zu einer Verlängerung der Lebensspanne bei den Sandgräbern beigetragen haben könnten: Zum einen wird die unterirdische Lebensweise genannt, von der man annimmt, dass sie im letzten gemeinsamen Vorfahren der Sandgräber entwickelt wurde und die von der Rohrratte nicht geteilt wird (Davies, et al. 2015). Diese schützende Umwelt habe möglicherweise zu einer drastischen Verringerung der extrinsischen Mortalität durch Fressfeinde geführt und so einen Selektionsvorteil für langlebige Individuen erzeugt (de Magalhaes, et al. 2007; Gorbunova, et al. 2014). Zum anderen wird angeführt, dass die beiden langlebigsten Sandgräbergattungen – Nacktmulle und Graumulle – eusozial sind (Jarvis and Bennett 1993). D.h., dass sie ähnlich wie staatenbildende Hautflügler (z.B. Bienen, Ameisen und Wespen) in Kolonien

zusammenleben, in denen sich nur wenige, bestimmte Individuen fortpflanzen (Costa and Fitzgerald 1996). Diese Individuen, häufig „Königin“ bzw. „König“ genannt, werden i.d.R. durch die reproduktiv nicht aktiven Mitglieder der Kolonie versorgt und beschützt, wodurch ihre extrinsische Mortalitätsrate gering ist und Langlebigkeit somit möglicherweise die evolutionäre Fitness begünstigt (Bennett and Faulkes 2000; Dammann, et al. 2011). Bei Insekten wurde gezeigt, dass die Evolution von Eusozialität mit einer 100-fachen Erhöhung der Lebensspanne assoziiert ist (Keller and Genoud 1997). Bei Säugetieren sind sich die Experten ebenfalls einig, dass Langlebigkeit positiv mit der der Evolution von Eusozialität korreliert, während umstritten ist, ob eine solche Beziehung auch zwischen Langlebigkeit und unterirdischer Lebensweise besteht (Healy, et al. 2014; Healy 2015; Williams and Shattuck 2015). Bisher haben sich drei Arbeiten mit der genomweiten Detektion positiver Selektion bei einzelnen Sandgräberarten bzw. bei ihrem letzten gemeinsamen Vorfahren beschäftigt (Kim, et al. 2011; Fang, Seim, et al. 2014; Davies, et al. 2015).

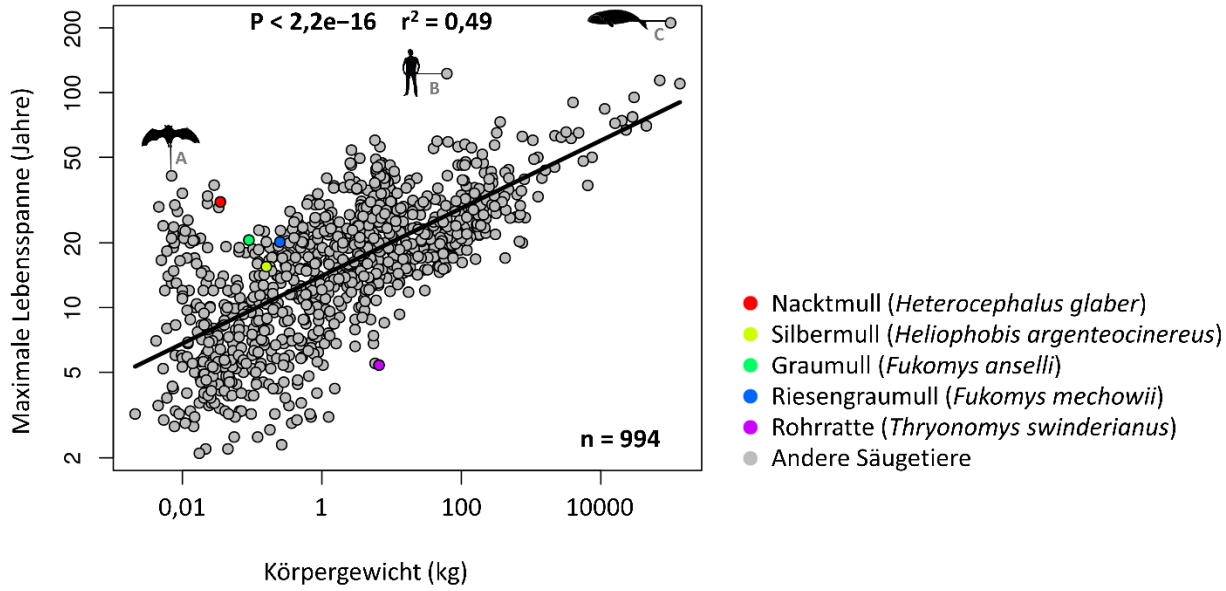



**Abbildung 5.** Beziehung zwischen Körpergewicht und maximaler Lebensspanne bei Säugern. Vier langlebige Sandgräberarten und der kurzlebige nächste Verwandte der Sandgräber – die Rohrratte – wurden farblich hervorgehoben (Legende rechts – neben Diagramm). Piktogramme markieren einige weitere, ausgewählte langlebige Spezies: A – Große Bartflederfledermaus (*Myotis brandtii*), B – Mensch (*Homo sapiens*), C – Grönlandwal (*Balaena mysticetus*). Die Zahl der eingetragenen Spezies (n) ist unten rechts im Diagramm angegeben. Die statistische Signifikanz (P) und das Bestimmtheitsmaß der Korrelation ($r^2$) werden am oberen Ende der Zeichenfläche angezeigt. Die zugrundeliegenden Daten stammen von AnAge (Tacutu, et al. 2013).

## 5. Die vorangegangenen Themen in den einzelnen Manuskripten

In den vorangegangenen Kapiteln wurde u.a. der Zweig-Positionstest zur Detektion positiver Selektion vorgestellt. Das von mir entwickelte und im ersten der folgenden Manuskripte beschriebene Programm PosiGene baut auf diesem Test auf und ermöglicht das automatisierte und genomweite Identifizieren von PSGs. In Manuskript I zeige ich insbesondere, dass die in PosiGene implementierten Filterstrategien geeignet sind, um trotz der bekannten Anfälligkeit des Zweig-Positionstest gegenüber Alignierungsproblemen, eine niedrige Falsch-Positiv-Rate und somit verlässliche PSG-Vorhersagen zu gewährleisten. PosiGene habe ich in den Manuskripten II, IV und V angewendet, um PSGs auf evolutionären Zweigen zu ermitteln, auf denen sich die Lebensspanne sehr wahrscheinlich verändert hat, und die Ergebnisse anschließend aus dem Blickwinkel der Alternsforschung interpretiert. In den Manuskripten II und IV habe ich den Zweig des Türkisen Prachtgrundkärpflings bzw. anzestrale Prachtgrundkärpflingzweige untersucht, die zeitlich mit historischen Klimaveränderungen

korrespondieren, von denen man annimmt, dass sie zur Verkürzung der Lebensspanne geführt haben. Manuskript III vergleicht die Ergebnisse zur positiven Selektion aus Manuskript II mit denen einer zeitgleich erschienenen Arbeit (Valenzano, et al. 2015) anhand für die Alternsforschung relevanter PSG-Vorhersagen. Mit Manuskript V untersuche ich positive Selektion auf Zweigen von existierenden und anzestralen Sandgräberarten, auf denen die Lebensspanne wahrscheinlich verlängert wurde.

# Übersicht der Manuskripte

| | | Artikel | Zeitschrift; IF*; Status | Eigenanteil |
|---|---|---|---|---|
| I | Titel | PosiGene: automated and easy-to-use pipeline for genome-wide detection of positively selected genes. | Nucleic Acids Research; IF: 9,2; **veröffentlicht** | 60% |
| | Autoren | **Arne Sahm**, Martin Bens, Matthias Platzer, Karol Szafranski | | |
| | Zusammenfassung | Es wurde ein Computerprogramm zur genomweiten Detektion positiver Selektion entwickelt. Das Hauptziel dabei war die Senkung der Falsch-Positiv-Rate durch geeignete Filterschritte. Die entsprechenden Nachweise wurden durch Simulationen und Tests auf Realdaten erbracht. | | |
| II | Titel | Insights into Sex Chromosome Evolution and Aging from the Genome of a Short-Lived Fish | Cell; IF: 28,7; **veröffentlicht** | 5% |
| | Autoren | Kathrin Reichwald, Andreas Petzold, Philipp Koch, Bryan R. Downie, Nils Hartmann, Stefan Pietsch, Mario Baumgart, Domitille Chalopin, Marius Felder, Martin Bens, **Arne Sahm**, Karol Szafranski, Stefan Taudien, Marco Groth, Ivan Arisi, Anja Weise, Samarth S. Bhatt, Virag Sharma, Johann M. Kraus, Florian Schmid, Steffen Priebe, Thomas Liehr, Matthias Görlach, Manuel E. Than, Michael Hiller, Hans A. Kestler, Jean-Nicolas Volff, Manfred Schartl, Alessandro Cellerino, Christoph Englert, Matthias Platzer | | |
| | Zusammenfassung | Der Artikel beschäftigt sich mit der Sequenzierung und den Eigenschaften des Genoms des Türkisenen Prachtgrundkärpflings – der kurzlebigsten Spezies, die zurzeit in Gefangenschaft gehalten werden kann. Ich habe dabei die Analysen zur positiven Selektion beigesteuert. | | |
| III | Titel | Outgroups and Positive Selection: The Nothobranchius furzeri Case | Trends in Genetics; IF: 9,9; **veröffentlicht** | 50% |
| | Autoren | **Arne Sahm**, Matthias Platzer, Alessandro Cellerino | | |
| | Zusammenfassung | In diesem Übersichtsartikel wurden die Ergebnisse der vorgenannten Arbeit und die der zeitgleich erschienenen Studie einer anderen Arbeitsgruppe miteinander verglichen. Dabei wurde insbesondere der Hauptgrund für die gefundenen Unterschiede aufgezeigt. | | |
| IV | Titel | Parallel evolution of genes controlling mitonuclear balance in short-lived annual fishes. | Aging Cell; IF: 5,8; **veröffentlicht** | 40% |
| | Autoren | **Arne Sahm**, Martin Bens, Matthias Platzer, Alessandro Cellerino | | |
| | Zusammenfassung | Der Artikel veröffentlicht Ergebnisse zur positiven Selektion auf anzestralen Zweigen der Prachtgrundkärpflinge, auf denen sehr wahrscheinlich die Lebenspanne verkürzt wurde. Dabei wurden insbesondere PSGs auf allen Ebenen der mitochondrialen Biogenese gefunden. | | |
| V | Titel | Long-lived rodents reveal signatures of positive selection in genes associated with lifespan and eusociality | Molecular Biology and Evolution; IF: 13,6; **eingereicht** | 30% |
| | Autoren | Arne Sahm, Martin Bens, Karol Szafranski, Susanne Holtze, Marco Groth, Matthias Görlach, Cornelis Calkhoven, Christine Müller, Matthias Schwab, Hans A. Kestler, Alessandro Cellerino, Hynek Burda, Thomas Hildebrandt, Philip Dammann, Matthias Platzer | | |
| | Zusammenfassung | Der Artikel veröffentlicht Ergebnisse zur positiven Selektion bei langlebigen Sandgräbern (Nacktmull, Graumulle etc.) und ihren letzten gemeinsamen Vorfahren. Dabei wurden viele alternsrelevante Gene und biologische Prozesse identifiziert, insbesondere solche, die vom mTOR-Signalweg reguliert werden. | | |

*IF – englisch: *Impact factor* (Einflussfaktor) 2015/16

# Manuskripte

## Manuskript I: PosiGene: automated and easy-to-use pipeline for genome-wide detection of positively selected genes.

*Nucleic Acids Research, 2017* **1**
*doi: 10.1093/nar/gkx179*

# PosiGene: automated and easy-to-use pipeline for genome-wide detection of positively selected genes

**Arne Sahm*, Martin Bens, Matthias Platzer and Karol Szafranski**

Leibniz Institute on Aging, Fritz Lipmann Institute, 07745 Jena, Germany

Received November 29, 2016; Revised March 03, 2017; Editorial Decision March 06, 2017; Accepted March 09, 2017

## ABSTRACT

**Many comparative genomics studies aim to find the genetic basis of species-specific phenotypic traits. A prevailing strategy is to search genome-wide for genes that evolved under positive selection based on the non-synonymous to synonymous substitution ratio. However, incongruent results largely due to high false positive rates indicate the need for standardization of quality criteria and software tools. Main challenges are the ortholog and isoform assignment, the high sensitivity of the statistical models to alignment errors and the imperative to parallelize large parts of the software. We developed the software tool PosiGene that (i) detects positively selected genes (PSGs) on genome-scale, (ii) allows analysis of specific evolutionary branches, (iii) can be used in arbitrary species contexts and (iv) offers visualization of the results for further manual validation and biological interpretation. We exemplify PosiGene's performance using simulated and real data. In the simulated data approach, we determined a false positive rate <1%. With real data, we found that 68.4% of the PSGs detected by PosiGene, were shared by at least one previous study that used the same set of species. PosiGene is a user-friendly, reliable tool for reproducible genome-wide identification of PSGs and freely available at https://github.com/gengit/PosiGene.**

## INTRODUCTION

'What is the genetic basis of phenotypic differences between species?' is a recurring question in comparative genomics. A frequently used method is to search for genes that evolved under positive selection. Positive selection describes the phenomenon that beneficial gene variants become fixed in a population/species over time because they increase fitness. It is a major evolutionary mechanism that leads to fixation of innovation and adaptation to changing environmental conditions (1,2). Most commonly, the **ω** ratio (the non-synonymous to synonymous substitution rate ratio, also known as $d_N/d_S$ or $K_a/K_s$) is used as a sign for positive selection on protein-coding genes.

Systematic scans for positively selected genes (PSGs) have provided insights into adaptation processes. For example, PSGs were identified for many well known bacterial pathogens that have immune related counterparts on the mammalian side (3–7). A similar 'arms race' can be found between venomous animals and their predators or preys (8,9). Genome-scale searches linked PSGs to phenotypic traits like subterranean life and longevity of mole-rats (10–12), the ability of Tibetan antelopes to live in high altitudes with low oxygen-concentration (13) and increased mitochondrial efficiency leading to lower ROS-levels in ants as potential prerequisite for their remarkable long lifespan (14). Moreover, a significant role of positive selection on neuronal-expressed genes in the evolution of the human nervous system was illustrated (15).

Despite important insights gained by many genome-wide works, re-evaluation studies have stated false-positive rates of predicted PSGs between 45 and 90% (16–19). As the respective original studies are based on locally developed and implemented computational tools, this led to heterogeneous quality standards, absence of reproducibility and eventually, to incongruent results (10,16,20).

There is a lack of a general software solution that offers automated and reliable analysis of genome-scale data. Several challenges are contributing to this situation. First, such a software solution must be applicable in a general way, which means that an ortholog assignment approach is required that allows arbitrary species sets to be used and consequently, arbitrary evolutionary branches to be tested. Second, the management of alternative splice variants is an important aspect in a eukaryotic context. Since the majority of eukaryotic genes are expressed as multiple transcripts it is necessary to select representative isoforms for further downstream analyzes. Choosing the longest isoform or picking at random can be a substantial source of false positives, because these approaches increase the chance of misalignments due to the inclusion of non-homologous regions, such as those derived from species-specific exons. In-

*To whom correspondence should be addressed. Tel: +49 3641 656050; Fax: +49 3641 656255; Email: arne.sahm@leibniz-fli.de
Present address: Arne Sahm, Genome Analysis, Leibniz Institute on Aging, Fritz Lipmann Institute, Jena, Thuringia, 07745, Germany.

**Table 1.** Features of existing software in the field of PSG identification

| Tasks/Challenges | Datamonkey[1] | Selecton[2] | JCoDA[3] | IDEA[4] | PhyleasProg[5] | PSP[6] | POTION[7] |
|---|---|---|---|---|---|---|---|
| Detection of ortholog relationships between genes of different species | - | - | - | - | + | + | ± |
| Calculation of coding sequence alignments | - | + | + | - | + | + | + |
| Reconstruction of a phylogenetic tree | + | + | + | + | + | + | + |
| Possibility to use multiple CPU cores to reduce running time | - | - | - | + | + | + | + |
| Filter steps to eliminate alignment errors and problematic results | ± | - | - | - | ± | + | + |
| Possibility to scan for PSGs within any user-defined species set | + | + | + | + | - | - | + |
| Option to test positive selection along specific branches | + | - | - | + | + | + | - |
| Visualization of the positively selected amino acids within the alignment | + | + | + | + | - | - | + |

[1] (23),[2] (24),[3] (25),[4] (26),[5] (27),[6] (28),[7] (29)

stead, isoforms should be chosen that are likely to be similar from a functional and evolutionary perspective – but also in a reasonable amount of time (21). Third, evolutionary codon models as the backbone of PSG identification are highly sensitive to bad quality of input data. Errors can originate from sequencing, assembly and gene annotation as well as pseudogenes that were not recognized as those. Furthermore, errors can occur during the different steps of the genome-wide PSG search itself, e.g., if gene fragments or poorly conserved sequences are assigned to an ortholog group. Another source of errors lies in applying the statistical models on alignments showing non-conserved regions that cannot be resolved without ambiguity. All these problems can lead to alignments of non-homologous codons resulting in a statistical signal that is misinterpreted as positive selection. On genome scale even low rates of false signals can outnumber the true candidates (16–18). This is why strict quality-filtering strategies are necessary to ensure reliable results (16,19). Fourth, it is imperative to efficiently parallelize large parts of the software, because most of the steps it has to conduct, like ortholog assignment, high quality multiple sequence alignment (MSA) and application of codon substitution models, have considerable computational costs. Execution of such steps on a single processor for thousands of genes is not practicable within a reasonable amount of time (22).

Genome-scale PSG searches require considerable experience in bioinformatics. To simplify the PSG search several attempts have been made over the recent years (Table 1). The tools Datamonkey (23), Selecton (24) and JCoDA (25) were developed to simplify the procedures for single-gene studies in particular steps: alignment of orthologous sequences, computation of the phylogenetic tree and/or configuration of tools that implement codon substitution models. IDEA (26) is a graphical program that allows to analyze multiple genes in parallel but requires pre-aligned sequence data and virtually lacks a filtering procedure or data quality control to ensure plausibility of the predicted candidates. PhyleasProg (27) and PSP (28) are able to perform all necessary steps for genome-wide PSG identification but are restricted to fix sets of few vertebrate species or bacteria strains, respectively. The recently developed end-to-end pipeline POTION (29) meets most of the requirements. However, it does not offer a solution for branch-specific PSG search, which is the common application scenario because it allows to link identified PSGs to phenotypic traits (1,10,14,15,30–35).

Toward user-friendly, reliable tools for reproducible genome-wide identification we developed PosiGene that addresses all the above mentioned challenges and performs the complex analysis automatically. In addition, PosiGene offers alignment visualization, in which positively selected protein sites and functional domains are highlighted. We validated PosiGene on simulated data using sequences with known features of positive selection and on real data comparing its results against those of five high-ranking publications on positive selection along the human lineage.

## MATERIALS AND METHODS

### Structure and workflow of the PosiGene pipeline

*Overview.* The minimal required input comprises coding sequences – in FASTA or GENBANK format – for all species to be analyzed. The output consists of a table showing all genes (including those that are not significant) ranked by their probability to be under positive selection and includes information about positively selected sites, $d_N/d_S$ ratios as well as links to alignment visualizations.

A user manual (https://github.com/gengit/PosiGene/blob/master/doc/User_Guide.pdf) provides detailed information about all possible parameters that can be used to customize PosiGene. The software is divided in three consecutive modules: the first module (M1) builds the ortholog catalog, i.e. the genome-wide set of ortholog assignments, based on the user-defined set of species and sequences. The second module (M2) constructs alignments and derives a phylogenetic species tree. The third module (M3) scans genes for positive selection along a user-chosen

branch of the species tree. PosiGene can be called in a way that all modules are executed consecutively or to run a single module separately. The latter feature can be used to add a species to the ortholog catalog, change parameters or to search another branch for PSGs without having to rerun the whole pipeline (Figure 1).

PosiGene is implemented in Perl and uses different Bioperl (36) modules for reading and writing sequence, tree as well as alignment files. All modules – except the HomoloGene based ortholog assignment at the beginning of M1 (see below), which stresses Input/Output – are heavily parallelized. Threads are created once at the beginning of each submodule (Figure 1) and are reused efficiently for new tasks by the main thread via queues. This avoids extra or inhomogeneous computational load caused by thread administration.

All arguments used by PosiGene to call incorporated third party programs are listed in Supplementary Table S1.

*M1: building the ortholog catalog.* The assignment of genes to ortholog groups is the basis of later analyses. We have implemented a mixture of core species with already established ortholog relations and automated orthology prediction for any user-supplied species' data. This ensures reliability as well as flexibility of the ortholog assignment system.

Ortholog groups are determined, in a first approach, based on the HomoloGene database (37). The local HomoloGene copy is contained in the program package and currently contains 21 species covering a wide evolutionary range (http://www.ncbi.nlm.nih.gov/homologene/statistics/). Sequences of species that are not part of HomoloGene are assigned to the initial ortholog groups by a best-bidirectional BLAST hit criterion (38,39), which was adapted to resolve multiple isoforms per gene, using group-to-group instead of sequence-to-sequence assignment. We define group-to-group assignment such that a gene $\boldsymbol{X}$ of a species that is not part of HomoloGene is assigned to a homology group $\boldsymbol{Y}$, as defined by HomoloGene, if and only if the best hit across all isoforms of $\boldsymbol{X}$ is within $\boldsymbol{Y}$ and vice versa. The best-bidirectional hit criterion was shown to perform well in comparison with other ortholog assignment methods, irrespective of phylogenetic distance (40).

The module M1 is skipped if the user provides ortholog assignments of the sequences.

*M2: alignments and phylogeny.* The first step in this module is a similarity-based sequence selection to ensure that, per subsequently conducted positive selection test, there will be only one transcript isoform per species. Therefore, to each isoform of an 'anchor species' the most similar isoform of each other species is assigned. The anchor species of a PosiGene run is chosen by the user and could be, as a recommendation, the best annotated species with the most complete set of coding sequences or a species whose lineage shall be tested subsequently for positive selection. The isoforms that are most similar to the anchor species' sequences are determined via an initial MSA on protein level calculated by CLUSTALW. For this all possible isoforms from each species in an ortholog group are used. In comparison to pure pairwise alignments, the progressive nature of CLUSTALW, which aligns more similar sequences first, decreases the chance of aligning non-homologous regions, such as alternative exons. In comparison to the subsequently used PRANK, the widely used aligner CLUSTALW is much faster and thus, be able to produce results on large, i.e. many sequence containing, MSAs in a feasible amount of time (41). This is important because many genes are spliced into multiple isoforms. Finally, there are as many isoform assignments per ortholog group as there are isoforms in the anchor species. Generally, all following procedures, including M3, will be applied to the obtained isoform assignments.

Next, highly divergent sequences are removed from the isoform assignments. Each non-anchor species sequence whose similarity with the anchor species sequence does not reach a threshold will be removed. Furthermore, in order to guarantee an adequate level of conservation between the non-anchor species sequences, each of them is required to fulfill a second similarity threshold, in respect to all other non-anchor species sequences. The latter rule is implemented by iteratively removing sequences, beginning with the sequence that violates the rule most often. If multiple sequences violate the rule with equal frequency, the sequence that has the lowest similarity to the anchor species sequence is removed first.

For subsequent analysis steps, a phylogenetic tree is needed. The user can either provide a species tree, or it will be computed from the previously calculated isoform assignments using the parsimony method of the PHYLIP package (42) and jackknifing. Briefly, for this step, those isoform assignments are used that contain, after aforementioned sequence filtering, still all species that were specified by the user at the beginning. The aligned isoform assignments are concatenated and then cut in chunks of equal length. Each chunk is filtered with GBLOCKS (43) to remove gaps and unreliable alignment columns, following a tree reconstruction based on the filtered chunks with DNAPARS of the PHYLIP package. Dnapars carries out unrooted parsimony (44) and uses the method of (45) to calculate branch lengths. From these trees a consensus tree is calculated with CONSENSE of the same package and unrooted afterward. Since CONSENSE does not predict consensus branch lengths, we calculate the average branch length for every node of the consensus tree over all nodes of the chunk trees that are equivalents of the respective consensus tree node.

All isoform assignments that comprise at least three sequences (which means also three species) are aligned now on codon level using PRANK (46). The choice of the alignment software has a large impact on the result of PSG identification (18,47). PRANK produces the most reliable candidates in this context, as was found on simulated as well as real data (18,19,48,49). As guide tree the species tree is used (see above).

*M3: positive selection and filtering.* To identify genes under positive selection on specific evolutionary branches, we use the PAML package (50,51). PAML is widely used as a framework to test phylogenetic hypotheses by using maximum likelihood based on estimation of the $\boldsymbol{\omega}$ ratio. Specifically, we use the CODEML program of the PAML package

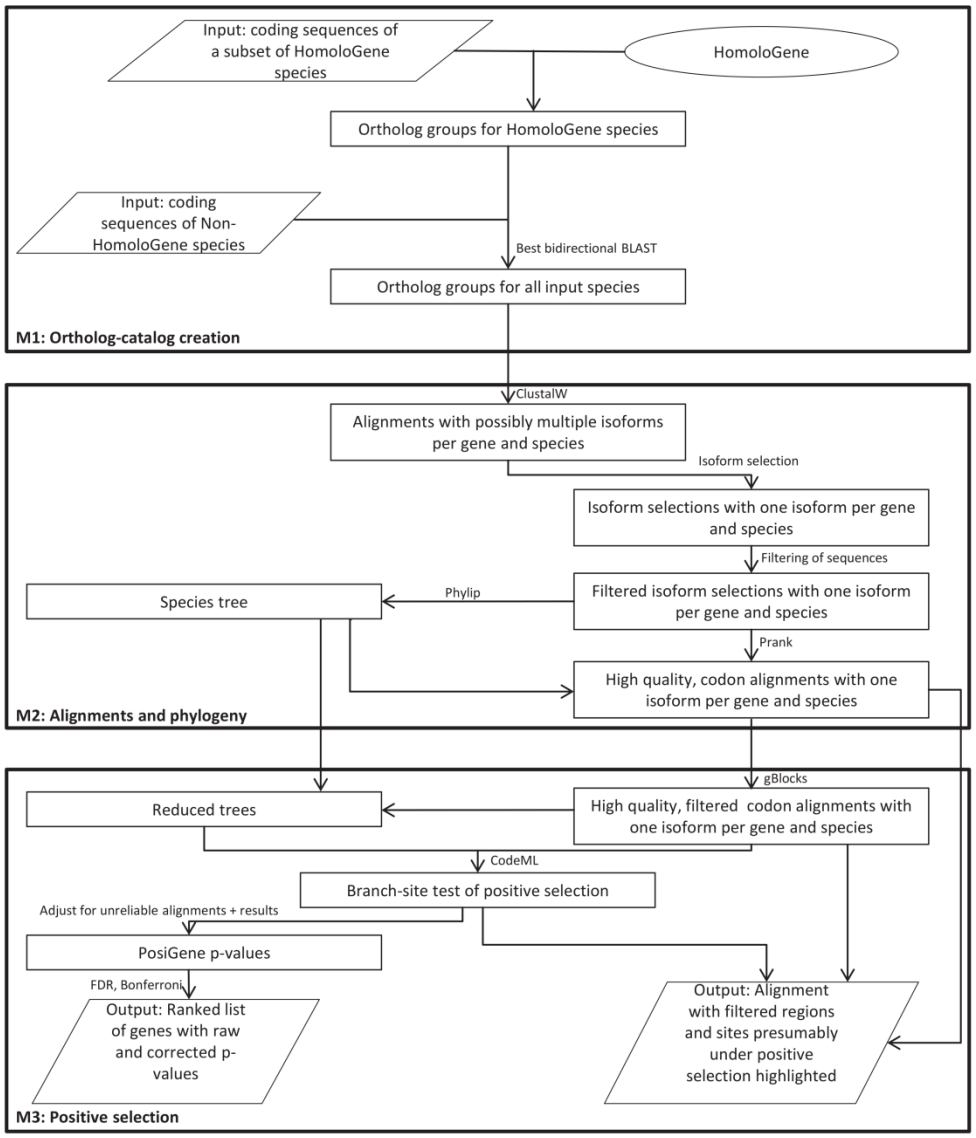



**Figure 1.** Workflow diagram of PosiGene.

to conduct the branch-site test of positive selection on each PRANK MSA (52,53). Briefly, this test is conducted by calculating and comparing the likelihoods of a null model, under which all sites may evolve under neutral or negative selection and an alternative model, under which the sites of the targeted branch are additionally allowed to evolve under positive selection. The *P*-value for the likelihood ratio test is calculated via a $\chi^2$ distribution with one degree of freedom. Besides a PRANK alignment, CODEML is supplied with a phylogenetic tree reduced to the species that are represented in the respective MSA, if necessary. Simulations have shown that the branch-site test has good accuracy and statistical power. However, it is sensitive to alignment as well as sequence errors and tends to produce more false negatives in scenarios of few, very similar or very short sequences due to low information content (54,55). Besides nominal *P*-values PosiGene results provide correction for multiple testing using the Bonferroni and Benjamini–Hochberg methods. Specific sites under positive selection are inferred by the Bayes emiprical Bayes method (56) implemented in CODEML.

As part of the PosiGene workflow, we paid special attention to minimize potential false positive PSGs by implementing a series of filtering steps (Figure 2). First, gaps and surrounding unconserved alignment columns are stringently removed with GBLOCKS (43) from the PRANK MSAs. A filtering of questionable alignment columns is necessary, because alignment of non-homologous codons is a major source of false positives (16). Second, as was mentioned, sequences failing pairwise similarity thresholds are deleted from alignments early in the workflow. MSAs con-

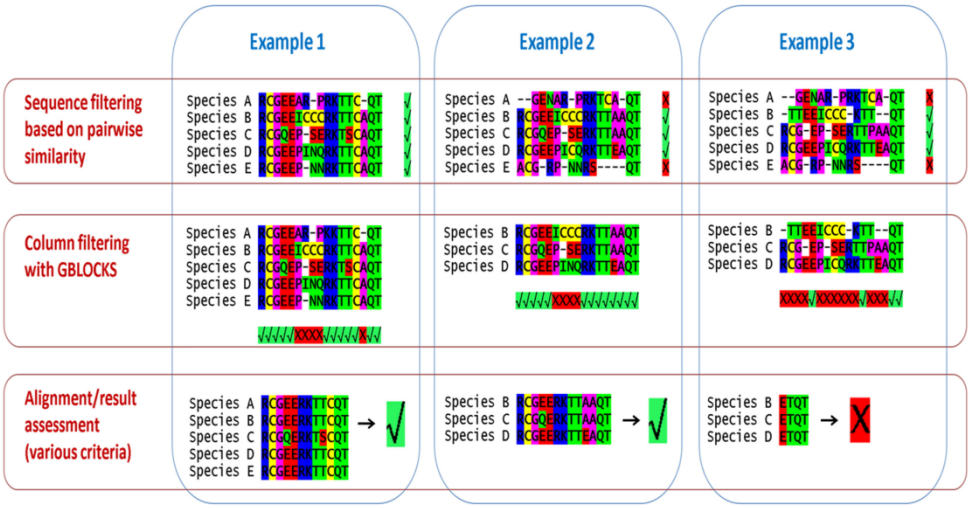



**Figure 2.** Schematic illustration of the filtering in PosiGene. Three approaches that are conducted at different steps of the program are depicted. The red marked 'X' means that the respective sequence/species, alignment column or the whole alignment/result were removed from further analysis, while the green marked √ means that the filter was passed. The shown examples are artificial and serve for demonstration only. In particular, for Example 3, the minimum length for a block of accepted alignment columns is depicted shorter (one codon/amino-acid) than in real application. The reason why the alignment in Example 3 does not pass the filter would be that a too small fraction of the alignment passed the column filter. For compact illustration, all steps are shown on protein level, while the column filtering works in reality on codon level.

taining those sequences are likely to have many disordered regions, promoting the alignment of non-homologous codons. This filtering step can also be seen as an instrument to reduce false negatives. Few badly conserved sequences can force the first mentioned filter to delete large parts of the MSA reducing the power of the test and potentially removing positively selected sites. Third, entire MSAs can be discarded if they are considered unreliable for the following reasons, if: (i) a small absolute number or a small percentage of alignment columns or anchor species codons remain after the first filtering step, (ii) few sequences remain after the second filtering step, (iii) disproportional $d_N/d_S$ ratios (e.g. ≥100 in foreground branch) were calculated by CODEML or (iv) an implausibly high fraction of positively selected sites was inferred. Additionally, MSAs will only be considered if at least one species from the sister taxon (i.e. the most closely-related species/clade) of the examined branch is represented in it. Without this condition it is not possible to say whether potentially observed selective pressure worked on the branch of interest or before in evolution (57).

The alignment visualization component processes four kinds of information: the MSA itself, the probability for each site to be under positive selection, which parts of the MSA were removed by GBLOCKS and thus could not be analyzed, as well as functional domains that are potentially listed in the GENBANK file of the anchor species. The information is depicted in two ways: first, as Portable Network Graphics (PNG) in different display formats based on Bioperl and the GD Graphics library; second, as a file type that is interpretable by Jalview (58). Jalview is a free Java based program for MSA visualization that is delivered with the PosiGene package and integrated insofar as PosiGene's Jalview visualizations can be opened with one simple command. Jalview also allows the user to edit the alignment, e.g., by adding further annotations.

### Validation methods

*Valdiation on simulated data.* First, we tested PosiGene based on simulated coding sequences that were generated with INDELible (59). Note, that the branch-site tests evaluates, for a given coding sequence, whether the assumption that a proportion of codons is target to positive selection on the tested branch fits the data significantly better than the assumption that all its codons evolved under neutral or purifying selection. Selective pressure is represented by the $\omega$ ratio and $\omega > 1$ indicates positive selection.

In order to assess the false positive rate we simulated the evolution of 1000 coding sequences by a selection scheme *N* without sites under positive selection. In scheme *N*, the sitewise selective pressure was set to a discrete distribution that was previously estimated based on 6.05 million codons in 12 871 gene trees comprising 29 mammals (60). However, we replaced the 0.99–1.0 quantile (the only one with $\omega > 1$) with the weighted average of all other quantiles $\omega = 0.21222$ (Supplementary Table S2). Indels were modeled with a geometric length distribution with parameter $q = 1-p = 0.35$ resulting in a mean and standard deviation of 1.54 respectively 0.91 codons. This distribution, developed in a similar simulation study (48), adequately fits published data on coding sequences of mammalian genomes (61,62). We used a ratio of substitution to indels of 43 as it was found in coding regions of primates (62). The ratio of transition to transversion substitutions, $\kappa$, was fixed at 2 and the stationary codon frequency of α-globin from our real data validation was used. For a realistic test scenario the sequences were evolved along the phylogenetic tree of the real data validation. However, the branch lengths had

to be multiplied with three in order to conform with a different concept of branch lengths used by INDELible. We verified that the branch lengths that were predicted by PosiGene on the simulated datasets match those of the original tree. All branches of the tree were simulated to evolve under selection scheme ***N*** (Supplementary Table S2). The root sequence length was set to 400. Finally, we configured PosiGene to search separately on one internal as well as on a terminal branch of the tree for PSGs to test the program for both possibilities. The tested terminal branch was the one that corresponds with the human branch in the real data validation (see Figure 3) and the internal branch corresponds with the last common ancestor of human, chimp and gorilla (GHC). Of note, both tested branches were simulated (as all others) to evolve under selection scheme ***N***, i.e. without positive selection.

In order to assess sensitivity we used in principle the same simulation model with the modification that the two tested branches now were evolved under selection schemes ***A–E***. The other branches were still simulated under selection scheme ***N*** as before. The schemes ***A–E*** differ from ***N*** insofar as a proportion of sites with $\omega > 1$ was added. The signal for positive selection was concentrated on 1, 3, 5, 7 and 9% of the codons for the schemes ***A–E***, respectively (Supplementary Table S2). Its strength was adjusted to fit an overall average $\omega$ of 0.9 - indicating still for a moderate purifying selection over the whole sequence. For each scheme ***A–E*** and for each of the both tested branches again 1000 sequences were generated.

*Validation on real data.* To determine the congruency among the five human studies as well as POTION and PosiGene results, we converted all candidate IDs to Ensembl human gene IDs. Due to historical reasons, multiple Ensembl gene IDs can refer to the same gene. Therefore we performed a last translation step and took the Ensembl gene names as objects of comparison. In congruency with most of the regarded studies (15,30,34,35) we defined candidates by having passed the filters of the respective work and nominal *P*-values equal or below 0.05 based on the branch-site test of positive selection. For ID conversion (Supplementary Table S3) we used Ensembl Biomart (63), except for the conversion of UCSC transcript IDs used by (1) to RefSeq transcript IDs for which we used the UCSC Table Browser (64). Additionally, for the OrthoMCL (65) cluster names that are used in the POTION output we determined the human protein IDs within the respective cluster and used them for further conversion. PosiGene were run on two different species sets: one with four species and one with nine species. Since there is no gold standard for PSGs, we define true candidates as being identified by at least two (respectively at least three) of the examined studies. Thus, the precision of a given study is defined as following:

$$\text{precision} = \frac{|\{\text{study candidates}\} \cap \{\text{true candidates}\}|}{|\{\text{study candidates}\}|}$$

## Benchmarking

For both PosiGene runs that were conducted in the frame of the real data validation, i.e. the four-species as well as the larger nine-species set, we measured how much total central processing unit (CPU) time was consumed and how much real time was needed to complete each of the three PosiGene modules (Table 2). For the benchmarking, we used a computer with 24 Intel Xeon processors of which each had a clock rate of 2.5 GHz. The differences between CPU time divided by the numbers of used processors and the real time that was needed, have to be mostly attributed to input/output operations on files. In the module M1 of the four-species run there is even less CPU time needed than real time due to the circumstance that all four-species were

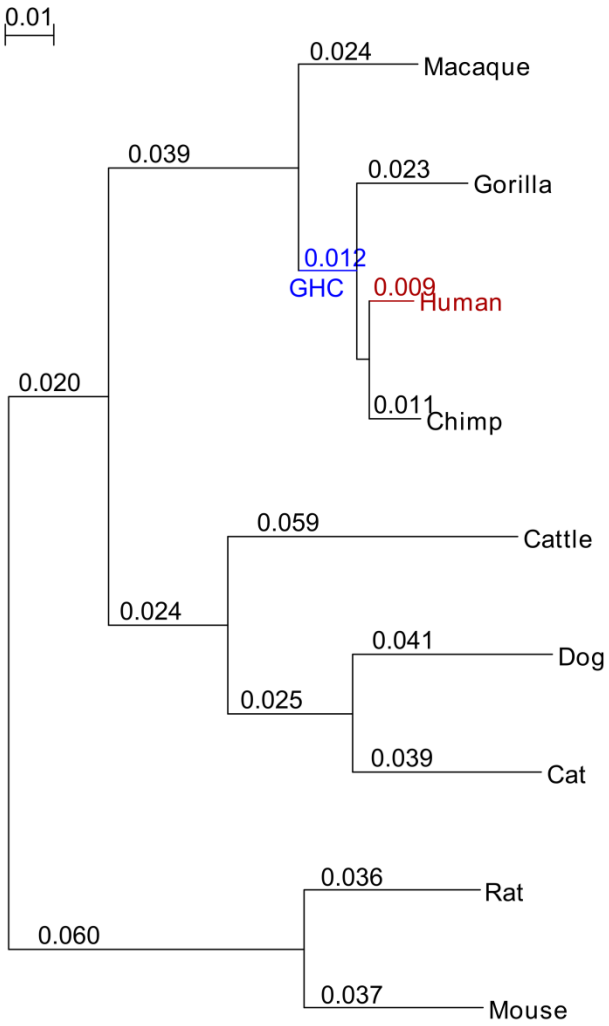



**Figure 3.** Species set used in real data validation. Shown is the phylogenetic tree that was calculated by PosiGene with the displayed species set. Branch lengths are drawn in scale and additionally shown directly at the branches. The respectively tested branches (both times human) are colored red. The tree was furthermore used to generate simulated sequences. In the simulation, the branches were tested that are equivalent to the red colored branch (human) and the blue colored branch (last common ancestor of GHC). In some of the simulation runs respectively one of these two branches was simulated to be under a different selection scheme than all other branches of the tree.

**Table 2.** Real and CPU time needed to run PosiGene on the real datasets analyzed in this work

| | 4-species set | | 9-species set | |
|---|---|---|---|---|
| | Real time | CPU time | Real time | CPU time |
| **M1: building the ortholog catalog** | 1.8 h | 1.1 h | 17.1 h | 305.3 h |
| **M2: alignments and phylogeny** | 6.6 h | 125.1 h | 26.8 h | 565.6 h |
| **M3: positive selection and filtering** | 5.1 h | 95.7 h | 33.9 h | 799.5 h |
| $\sum$ | 13.5 h | 221.9 h | 77.8 h | 1670.4 h |

Note: the table shows the real and CPU times consumed by two PosiGene runs that were executed on species sets of different sizes. A server with 24 processors was used for both runs.

part of the HomoloGene database and thus no BLAST steps were performed (see M1: building the ortholog catalog). PosiGene's memory consumption is negligible.

## RESULTS AND DISCUSSION

The newly developed end-to-end pipeline PosiGene is the first bioinformatics tool for the detection of PSG that performs the following analysis steps automatically: (i) determination of ortholog relationships between genes of different species, (ii) calculation of coding sequence alignments, (iii) reconstruction of a phylogenetic tree, (iv) filtering procedures for unreliable alignment data and implausible results as well as (v) the branch-site test of positive selection. Each step is heavily parallelized to reduce running time. PosiGene consists of three modules: M1 ortholog catalog creation, M2 alignments and phylogeny, M3 positive selection (Figure 1). It offers alignment visualization, in which positively selected protein sites and functional domains are highlighted. This enables biologists to manually validate and functionally interpret specific sites in individual candidates (Figure 4). Additionally, non-experts get an easy-to-use tool with reliable default parameters, while experts can configure the program to their needs and make use of its modularization. The PosiGene pipeline was applied successfully in several studies for genome-scale PSG identification (57,66,67). PosiGene is designed to run on linux platforms instantly without further installation and is available at https://github.com/gengit/PosiGene.

To validate PosiGene's performance we used simulated and real data.

### Validation on simulated data

First, we validated PosiGene on simulated coding sequences. The basic idea of this approach is to simulate the evolution of protein-coding sequences with defined selection schemes along the branches of a phylogenetic tree. This enabled us to create scenarios, in which PosiGene should detect positive selection (scenarios ***A–E***) and a scenario in which it should not (scenario ***N***). As tree we used the same as in the real data approach (Figure 3) and tested, in each scenario, the branches: (i) human, as a representative of a terminal branch or (ii) the last common ancestor of gorilla, human and chimp (GHC), as an internal branch (Table 3).

PosiGene results ($p \leq 0.05$) of scenario ***N*** indicate false positive rates of 0.3 and 0.4% in the human and the GHC branch, respectively. The true positive rates, determined in scenarios ***A–E***, lie between 5.4 and 30.7%, Supplementary Figure S1 shows false and true positive rates depending on how the *P*-value threshold is chosen. In order to assess PosiGene's false and true positive rates, we compared them with values from previously published extensive simulation experiments (48). In this study, Fletcher and Yang reported for the branch-site test of positive selection false positive rates without filtering between 2.1 and 13.0%. If, as only filtering procedure, gaps were removed from the alignment the false positive rates were between 2.4 and 10.2%. If alignment methods other than PRANK were used, the false positive rates were even higher. So, with a rate of 0.3–0.4% PosiGene's filtering of the alignments efficiently suppresses false positives. This, however, raises the question of whether our strict filtering procedures diminish PosiGene's true positive rate? This would be the case if the filtering removes true alignment signals. Since in simulations the 'true alignments' are known, Fletcher and Yang used these alignments directly to assess the maximum true positive rate that can technically be achieved using the branch-site test, and obtained rates between 1.4 and 33.1% (48). The fact that PosiGene's true positive rates (5.4–30.7%) were within the upper range of these estimates indicates that the negative impact of its filtering procedures is low. In regard to the still relatively high number of false negatives produced by the branch-site test, it should be noted that the coding sequences were simulated with an overall average signal of moderate negative selection and only a small fraction of codons were allowed to evolve under positive selection.

Furthermore, we observe that PosiGene's sensitivity is positively correlated with the concentration of the signal of positive selection (Pearson correlation; $r^2 = 0.89$, *P*-value 0.04), i.e. PosiGene's ability to detect positive selection increases if few sites are affected by heavy selective pressure (scenario ***A***) and decreases if many sites are influenced by weak selective pressure (scenarios ***B–E***).

### Validation on real data

While simulations offer the advantage of precise knowledge about the selective pressures that influenced sequence evolution, they may not cover the full range of problems that occur in analysis of real genome-wide data, e.g. the existence of paralogs, which currently cannot be simulated as above. However, since there is a lack of an independent validation technique for real data PSG candidates, we could not define a single study as a gold standard. Instead, we used the agreement between different studies as an indication for precision (positive predictive value) of predictions. Previous works have pinpointed precision in favor of sensitivity as major goal of PSG analysis on genome scale (16–18,20).

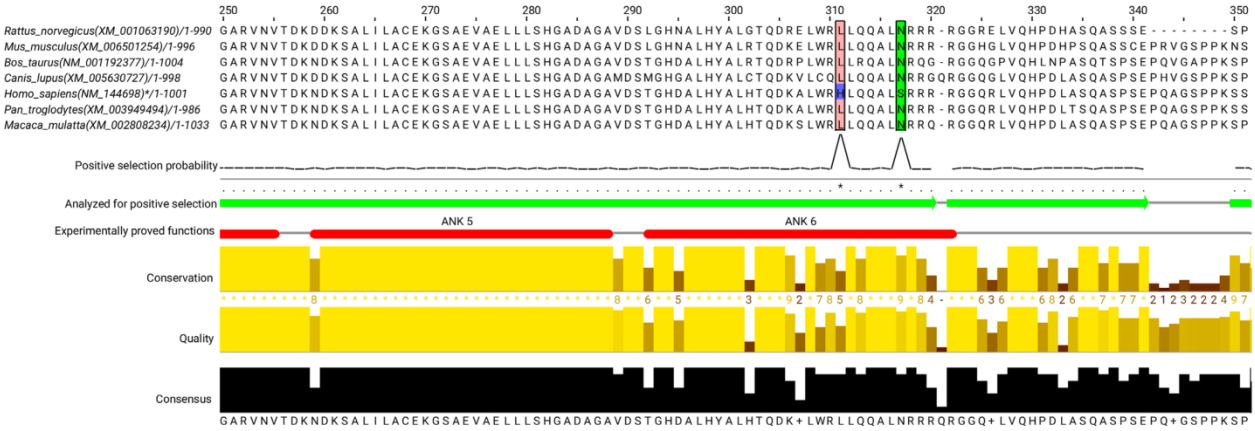



**Figure 4.** PSG visualization by PosiGene. Shown is a subregion of the ANKRD35 alignment. Human residues identified to be under positive selection (L311H, N317S) are colored with respect to physico-chemical properties using the Zappo code (http://www.jalview.org/help/html/colourSchemes/zappo.html). The probability of each residue to be positively selected is indicated by a line below the alignment and displayed upon a mouse-over action. Below are highlighted parts of the alignment that were used for the PSG test (green arrows) as well as experimentally supported protein domains based on an annotated sequence file (red bars). The three plots for conservation, quality and consensus at the bottom represent column-wise measures for the conservation of the physico-chemical properties of the amino-acids based on the Analysis of Multiply Aligned Sequences (AMAS) method (68), the likelihood of observing the mutations based on the BLOSUM62 matrix (69) and the percentage of the modal residue, respectively.

**Table 3.** PosiGene's performance on simulated gene trees

| Scenario | Codons under selection | Tested branch | Identified PSGs[1] | Description |
|---|---|---|---|---|
| ***N*** | 0% | Human | 0.3% | False positive rates |
| | | GHC[2] | 0.4% | |
| ***A*** | 1% | Human | 30.3% | True positive rates[3] |
| | | GHC | 30.7% | |
| ***B*** | 3% | Human | 15.0% | |
| | | GHC | 15.0% | |
| ***C*** | 5% | Human | 8.6% | |
| | | GHC | 10.2% | |
| ***D*** | 7% | Human | 6.2% | |
| | | GHC | 7.4% | |
| ***E*** | 9% | Human | 5.4% | |
| | | GHC | 6.1% | |

[1]A PSG was defined by having a nominal *P*-value $\leq 0.05$.
[2]GHC – last common ancestor of gorilla, human and chimp.
[3]The overall strength of positive selection was identical for scenarios *A*–*E* ($\omega = 0.9$) resulting in highest concentration of the selection pressure in scenario *A* and lowest in *E*.

For the real data validation approach, human served as a useful lineage because it has been analyzed multiple times for PSGs on a genome wide scale. Therefore, we took the PSG candidates from five human studies (1,15,31,33–35). In addition, we compared PosiGene only to the recently developed end-to-end pipeline POTION due to the principal limitations of other existing tools (Table 1). We ran POTION with default settings and complete mRNA sequence sets from human, chimp, mouse, rat, dog and maquaque as input (Supplementary Table S4). This species set is reduced in comparison to the set that was given to PosiGene due to the limitations of the OrthoMCL-based ortholog assignment system used by POTION, which restricts easy, semi-automatic ortholog assignment to species that are present in the OrthoMCL database (65). The species additionally used for the PosiGene run were cattle, cat and gorilla (Figure 3). To test the effect of the size of the used species set, an independent PosiGene run with only four species was conducted: human, chimp, maquaque and mouse. The PSG candidates of both PosiGene runs (Supplementary Tables S5 and 6) were predicted with default settings and human was set to be the tested species. Details about the examined studies like used alignment software, species set and filtering mechanisms are summarized in Supplementary Table S7.

We measured consensus on two levels: PSGs that were found by at least one, respectively, two other studies (Table 4, Supplementary Tables S8 and 9). The study of Clark *et al.* (33) shows least consistency with the other works. Since it is the earliest work, this could be explained by fewer and less qualitative gene sequences, availability of only two species for comparison to the tested human branch and use of an older version of the branch-site test that was improved subsequently (52,53). Also the POTION pipeline produced small intersections with the other works. However, this performance is hardly comparable because POTION uses site tests, which check whether a gene was generally under pos-

**Table 4.** Congruency of human PSG predictions across different studies with PosiGene nine-species result

| Study | Found PSGs | Shared by at least one other study | | Shared by at least two other studies | |
|---|---|---|---|---|---|
| | | Absolute | Precision [%] | Absolute | Precision [%] |
| Clark, *et al.* (33) | 525 | 22 | 4.2 | 9 | 1.7 |
| Arbiza, *et al.* (35) | 146 | 61 | 41.8 | 41 | 28.1 |
| Bakewell, *et al.* (34) | 138 | 88 | 63.8 | 56 | 40.6 |
| Kosiol, *et al.* (1) | 204 | 103 | 50.5 | 59 | 29.0 |
| Gaya-Vidal and Alba (15) | 190 | 65 | 34.2 | 43 | 22.7 |
| POTION | 123 | 8 | 6.5 | 5 | 4.1 |
| PosiGene | 98 | 67 | 68.4 | 47 | 48.0 |

itive selection during evolution, instead of the branch-site tests performed by the other works. The scope of application cases given with the presentation of POTION suggests that the program's default parameters, especially the filtering parameters, were optimized for PSG analysis in bacterial or less complex eukaryote genomes (29). Finally, we remark that the works of Kosiol (1) and Bakewell (34) show the best results in terms of sensitivity, that is, the absolute number of predicted PSGs confirmed by other studies.

On both measured consensus levels, PosiGene has consistently the highest precision, with more than two-third and almost the half of genes that were found by at least one, respectively, two other studies. This outperformance is not explained by the size of the species set used for branch-site analysis. A reduction of the species set from nine to four results in even a slightly increased precision, regarding PSGs that are shared by at least one other study and only in a minimal drop of precision from 48.0 to 44.7%, regarding PSGs that shared by at least two other studies (Supplementary Table S10). While the reduction of the species set does not negatively affect precision it does reduce sensitivity: the number of identified PSGs drops from 98 to 47. However, this is expected due to the decreased power of the branch-site test in alignments with fewer sequences (54). We acknowledge that, within the comparison, PosiGene identifies the fewest PSG candidates, potentially indicating a high false negative rate. This could be attributed to the circumstance that we laid our focus on precision instead of sensitivity, in agreement with the literature (16–18,20). In respect to co-supported candidates, however, only the Bakewell and Kosiol studies (1,34) identified more PSGs showing that PosiGene's sensitivity can compete with that of the other studies. Of note, the fully automated pipeline of PosiGene is compared against the primary results of high ranking studies, which were able to use tailored data quality controls that are difficult to implement in a generally applicable program. For example, the Bakewell study, which has the highest precision besides of PosiGene, integrated the nucleotide qualities of the chimpanzee genome as a main filtering mechanism into their approach. Furthermore, the studies neither had the aim nor provided tools to reproduce their approach. Arbiza, Bakewell and Gaya-Vidal (15,34,35) also did not provide the alignments which further hinders evaluation of the results and follow-up studies. In contrast, PosiGene offers the possibility of easy reproduction of results that were predicted by others and provides alignment visualizations to manually verify, biologically interpret and experimentally examine PSGs and selected sites.

## CONCLUSIONS

The identification of PSGs is a prevailing genomics approach that enabled insights into adaptation processes, molecular function and the genetic source of species-specific phenotypic traits. PosiGene can be used with a single command line call to search for relevant candidates on a user-chosen evolutionary branch and a genome-wide scale. Besides a list of genes that are ranked according to the probability to be under positive selection, PosiGene generates alignment visualizations which enable to contextually interpret the positively selected amino acid sites within the respective candidate.

We compared the functionality of PosiGene with other tools that partly enable to search for PSGs on different scales. We argue that none of them would be suited as a broadly applicable tool for genome-wide searches that aim to link phenotypic traits of a species or clade to its PSGs, because important aspects like filtering mechanisms, a freely selectable species set or a branch-specific analysis are lacking.

We demonstrated PosiGene's performance in two complementary validation strategies. One validation was based on simulated data giving precise control over targets of positive selection. It was shown that PosiGene's filter mechanisms result in a very small false positive rate that is a fraction of known values for unfiltered data. Since simulated data may not cover the full range of possible problems, a second validation on real data was performed. The results demonstrated that PosiGene reaches a good overlap with existing high-ranking studies on the human lineage, e.g., more than two-third of the PSGs that were identified by PosiGene were also found by at least one human study.

Altogether, we provide PosiGene as step toward a user-friendly tool for genome-wide identification of PSGs that produces reliable results reproducible by others which can be visualized for further manual validation and biological interpretation.

## AVAILABILITY

Project name: PosiGene
Project home page: https://github.com/gengit/PosiGene
Operating System: linux 64-Bit
Programming language: perl
License: GPL Version 3

## SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

## FUNDING

Leibniz-Gesellschaft through the Senatsausschusswettbewerb (SAW) [SAW-2012-FLI-2]; Deutsche Forschungsgemeinschaft (DFG) [PL 173/8-1].
*Conflict of interest statement.* None declared.

## REFERENCES

1. Kosiol,C., Vinar,T., da Fonseca,R.R., Hubisz,M.J., Bustamante,C.D., Nielsen,R. and Siepel,A. (2008) Patterns of positive selection in six mammalian genomes. *PLoS Genet.*, **4**, e1000144.
2. Yang,Z. (2005) The power of phylogenetic comparison in revealing protein function. *Proc. Natl. Acad. Sci. U.S.A*, **102**, 3179–3180.
3. Lefebure,T. and Stanhope,M.J. (2007) Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. *Genome Biol.*, **8**, R71.
4. Petersen,L., Bollback,J.P., Dimmic,M., Hubisz,M. and Nielsen,R. (2007) Genes under positive selection in Escherichia coli. *Genome Res.*, **17**, 1336–1343.
5. Soyer,Y., Orsi,R.H., Rodriguez-Rivera,L.D., Sun,Q. and Wiedmann,M. (2009) Genome wide evolutionary analyses reveal serotype specific patterns of positive selection in selected Salmonella serotypes. *BMC Evol. Biol.*, **9**, 264.
6. Suzuki,H., Lefebure,T., Bitar,P.P. and Stanhope,M.J. (2012) Comparative genomic analysis of the genus Staphylococcus including Staphylococcus aureus and its newly described sister species Staphylococcus simiae. *BMC Genomics*, **13**, 38.
7. Webb,A.E., Gerek,Z.N., Morgan,C.C., Walsh,T.A., Loscher,C.E., Edwards,S.V. and O'Connell,M.J. (2015) Adaptive evolution as a predictor of species-specific innate immune response. *Mol. Biol. Evol.*, **32**, 1717–1729.
8. Kozminsky-Atias,A. and Zilberberg,N. (2012) Molding the business end of neurotoxins by diversifying evolution. *FASEB J.*, **26**, 576–586.
9. Zhu,S., Bosmans,F. and Tytgat,J. (2004) Adaptive evolution of scorpion sodium channel toxins. *J. Mol. Evol.*, **58**, 145–153.
10. Davies,K.T., Bennett,N.C., Tsagkogeorga,G., Rossiter,S.J. and Faulkes,C.G. (2015) Family wide molecular adaptations to underground life in african mole-rats revealed by phylogenomic analysis. *Mol. Biol. Evol.*, **32**, 3089–3107.
11. Fang,X., Nevo,E., Han,L., Levanon,E.Y., Zhao,J., Avivi,A., Larkin,D., Jiang,X., Feranchuk,S., Zhu,Y. *et al.* (2014) Genome-wide adaptive complexes to underground stresses in blind mole rats Spalax. *Nat. Commun.*, **5**, 3966.
12. Fang,X., Seim,I., Huang,Z., Gerashchenko,M.V., Xiong,Z., Turanov,A.A., Zhu,Y., Lobanov,A.V., Fan,D., Yim,S.H. *et al.* (2014) Adaptations to a subterranean environment and longevity revealed by the analysis of mole rat genomes. *Cell Rep.*, **8**, 1354–1364.
13. Ge,R.L., Cai,Q., Shen,Y.Y., San,A., Ma,L., Zhang,Y., Yi,X., Chen,Y., Yang,L., Huang,Y. *et al.* (2013) Draft genome sequence of the tibetan antelope. *Nat. Commun.*, **4**, 1858.
14. Roux,J., Privman,E., Moretti,S., Daub,J.T., Robinson-Rechavi,M. and Keller,L. (2014) Patterns of positive selection in seven ant genomes. *Mol. Biol. Evol.*, **31**, 1661–1685.
15. Gaya-Vidal,M. and Alba,M.M. (2014) Uncovering adaptive evolution in the human lineage. *BMC Genomics*, **15**, 599.
16. Mallick,S., Gnerre,S., Muller,P. and Reich,D. (2009) The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res.*, **19**, 922–933.
17. Schneider,A., Souvorov,A., Sabath,N., Landan,G., Gonnet,G.H. and Graur,D. (2009) Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. *Genome Biol. Evol.*, **1**, 114–118.
18. Markova-Raina,P. and Petrov,D. (2011) High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. *Genome Res.*, **21**, 863–874.
19. Privman,E., Penn,O. and Pupko,T. (2012) Improving the performance of positive selection inference by filtering unreliable alignment regions. *Mol. Biol. Evol.*, **29**, 1–5.
20. Biswas,S. and Akey,J.M. (2006) Genomic insights into positive selection. *Trends Genet.*, **22**, 437–446.
21. Villanueva-Canas,J.L., Laurie,S. and Alba,M.M. (2013) Improving genome-wide scans of positive selection by using protein isoforms of similar length. *Genome Biol. Evol.*, **5**, 457–467.
22. Moretti,S., Murri,R., Maffioletti,S., Kuzniar,A., Castella,B., Salamin,N., Robinson-Rechavi,M. and Stockinger,H. (2012) gcodeml: a grid-enabled tool for detecting positive selection in biological evolution. *Stud. Health Technol. Inform.*, **175**, 59–68.
23. Delport,W., Poon,A.F., Frost,S.D. and Kosakovsky Pond,S.L. (2010) Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics*, **26**, 2455–2457.
24. Stern,A., Doron-Faigenboim,A., Erez,E., Martz,E., Bacharach,E. and Pupko,T. (2007) Selecton 2007: advanced models for detecting positive and purifying selection using a bayesian inference approach. *Nucleic Acids Res.*, **35**, W506–W511.
25. Steinway,S.N., Dannenfelser,R., Laucius,C.D., Hayes,J.E. and Nayak,S. (2010) JCoDA: a tool for detecting evolutionary selection. *BMC Bioinformatics*, **11**, 284.
26. Egan,A., Mahurkar,A., Crabtree,J., Badger,J.H., Carlton,J.M. and Silva,J.C. (2008) IDEA: interactive display for evolutionary analyses. *BMC Bioinformatics*, **9**, 524.
27. Busset,J., Cabau,C., Meslin,C. and Pascal,G. (2011) PhyleasProg: a user-oriented web server for wide evolutionary analyses. *Nucleic Acids Res.*, **39**, W479–W485.
28. Su,F., Ou,H.Y., Tao,F., Tang,H. and Xu,P. (2013) PSP: rapid identification of orthologous coding genes under positive selection across multiple closely related prokaryotic genomes. *BMC Genomics*, **14**, 924.
29. Hongo,J.A., de Castro,G.M., Cintra,L.C., Zerlotini,A. and Lobo,F.P. (2015) POTION: an end-to-end pipeline for positive Darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes. *BMC Genomics*, **16**, 567.
30. Clark,A.G., Eisen,M.B., Smith,D.R., Bergman,C.M., Oliver,B., Markow,T.A., Kaufman,T.C., Kellis,M., Gelbart,W., Iyer,V.N. *et al.* (2007) Evolution of genes and genomes on the Drosophila phylogeny. *Nature*, **450**, 203–218.
31. Gibbs,R.A., Rogers,J., Katze,M.G., Bumgarner,R., Weinstock,G.M., Mardis,E.R., Remington,K.A., Strausberg,R.L., Venter,J.C., Wilson,R.K. *et al.* (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science*, **316**, 222–234.
32. Parker,J., Tsagkogeorga,G., Cotton,J.A., Liu,Y., Provero,P., Stupka,E. and Rossiter,S.J. (2013) Genome-wide signatures of convergent evolution in echolocating mammals. *Nature*, **502**, 228–231.
33. Clark,A.G., Glanowski,S., Nielsen,R., Thomas,P.D., Kejariwal,A., Todd,M.A., Tanenbaum,D.M., Civello,D., Lu,F., Murphy,B. *et al.* (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science*, **302**, 1960–1963.
34. Bakewell,M.A., Shi,P. and Zhang,J. (2007) More genes underwent positive selection in chimpanzee evolution than in human evolution. *Proc. Natl. Acad. Sci. U.S.A*, **104**, 7489–7494.
35. Arbiza,L., Dopazo,J. and Dopazo,H. (2006) Positive selection, relaxation, and acceleration in the evolution of the human and chimp genome. *PLoS Comput. Biol.*, **2**, e38.
36. Stajich,J.E., Block,D., Boulez,K., Brenner,S.E., Chervitz,S.A., Dagdigian,C., Fuellen,G., Gilbert,J.G., Korf,I., Lapp,H. *et al.* (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–1618.
37. Geer,L.Y., Marchler-Bauer,A., Geer,R.C., Han,L., He,J., He,S., Liu,C., Shi,W. and Bryant,S.H. (2010) The NCBI BioSystems database. *Nucleic Acids Res.*, **38**, D492–D496.
38. Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl. Acad. Sci. U.S.A.*, **96**, 2896–2901.
39. Camacho,C., Coulouris,G., Avagyan,V., Ma,N., Papadopoulos,J., Bealer,K. and Madden,T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.
40. Altenhoff,A.M. and Dessimoz,C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput. Biol.*, **5**, e1000262.
41. Liu,K., Linder,C.R. and Warnow,T. (2010) Multiple sequence alignment: a major challenge to large-scale phylogenetics. *PLoS Curr.*, **2**, RRN1198.
42. Felsenstein,J. (2005) *PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author*, Department of Genome Sciences, University of Washington, Seattle.

43. Castresana,J. (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol. Biol. Evol.*, **17**, 540–552.
44. Eck,R.V. and Dayhoff,M.O. (1966) Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. *Science*, **152**, 363–366.
45. Hochbaum,D.S. and Pathria,A. (1997) Path costs in evolutionary tree reconstruction. *J. Comput. Biol.*, **4**, 163–175.
46. Loytynoja,A. and Goldman,N. (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science*, **320**, 1632–1635.
47. Wong,K.M., Suchard,M.A. and Huelsenbeck,J.P. (2008) Alignment uncertainty and genomic analysis. *Science*, **319**, 473–476.
48. Fletcher,W. and Yang,Z. (2010) The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol. Biol. Evol.*, **27**, 2257–2267.
49. Jordan,G. and Goldman,N. (2012) The effects of alignment error and alignment filtering on the sitewise detection of positive selection. *Mol. Biol. Evol.*, **29**, 1125–1139.
50. Yang,Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.
51. Yang,Z. (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.*, **13**, 555–556.
52. Yang,Z. and Nielsen,R. (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.*, **19**, 908–917.
53. Zhang,J., Nielsen,R. and Yang,Z. (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.*, **22**, 2472–2479.
54. Anisimova,M., Bielawski,J.P. and Yang,Z. (2001) Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.*, **18**, 1585–1592.
55. Yang,Z. and dos Reis,M. (2011) Statistical properties of the branch-site test of positive selection. *Mol. Biol. Evol.*, **28**, 1217–1228.
56. Yang,Z., Wong,W.S. and Nielsen,R. (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.*, **22**, 1107–1118.
57. Sahm,A., Platzer,M. and Cellerino,A. (2016) Outgroups and positive selection: the nothobranchius furzeri case. *Trends Genet.*, **32**, 523–525.
58. Waterhouse,A.M., Procter,J.B., Martin,D.M., Clamp,M. and Barton,G.J. (2009) Jalview version 2–a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, **25**, 1189–1191.
59. Fletcher,W. and Yang,Z. (2009) INDELible: a flexible simulator of biological sequence evolution. *Mol. Biol. Evol.*, **26**, 1879–1888.
60. Lindblad-Toh,K., Garber,M., Zuk,O., Lin,M.F., Parker,B.J., Washietl,S., Kheradpour,P., Ernst,J., Jordan,G., Mauceli,E. *et al.* (2011) A high-resolution map of human evolutionary constraint using 29 mammals. *Nature*, **478**, 476–482.
61. Taylor,M.S., Ponting,C.P. and Copley,R.R. (2004) Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. *Genome Res.*, **14**, 555–566.
62. Chen,J.Q., Wu,Y., Yang,H., Bergelson,J., Kreitman,M. and Tian,D. (2009) Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Mol. Biol. Evol.*, **26**, 1523–1531.
63. Kinsella,R.J., Kahari,A., Haider,S., Zamora,J., Proctor,G., Spudich,G., Almeida-King,J., Staines,D., Derwent,P., Kerhornou,A. *et al.* (2011) Ensembl biomarts: a hub for data retrieval across taxonomic space. *Database (Oxford)*, bar030.
64. Karolchik,D., Hinrichs,A.S., Furey,T.S., Roskin,K.M., Sugnet,C.W., Haussler,D. and Kent,W.J. (2004) The UCSC table browser data retrieval tool. *Nucleic Acids Res.*, **32**, D493–D496.
65. Chen,F., Mackey,A.J., Stoeckert,C.J. Jr and Roos,D.S. (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res.*, **34**, D363–368.
66. Reichwald,K., Petzold,A., Koch,P., Downie,B.R., Hartmann,N., Pietsch,S., Baumgart,M., Chalopin,D., Felder,M., Bens,M. *et al.* (2015) Insights into sex chromosome evolution and aging from the genome of a short-lived fish. *Cell*, **163**, 1527–1538.
67. Sahm,A., Bens,M., Platzer,M. and Cellerino,A. (2017) Parallel evolution of genes controlling mitonuclear balance in short-lived annual fishes. *Aging Cell*, doi:10.1111/acel.12577.
68. Livingstone,C.D. and Barton,G.J. (1993) Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation. *Comput. Appl. Biosci.*, **9**, 745–756.
69. Henikoff,S. and Henikoff,J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. U.S.A.*, **89**, 10915–10919.

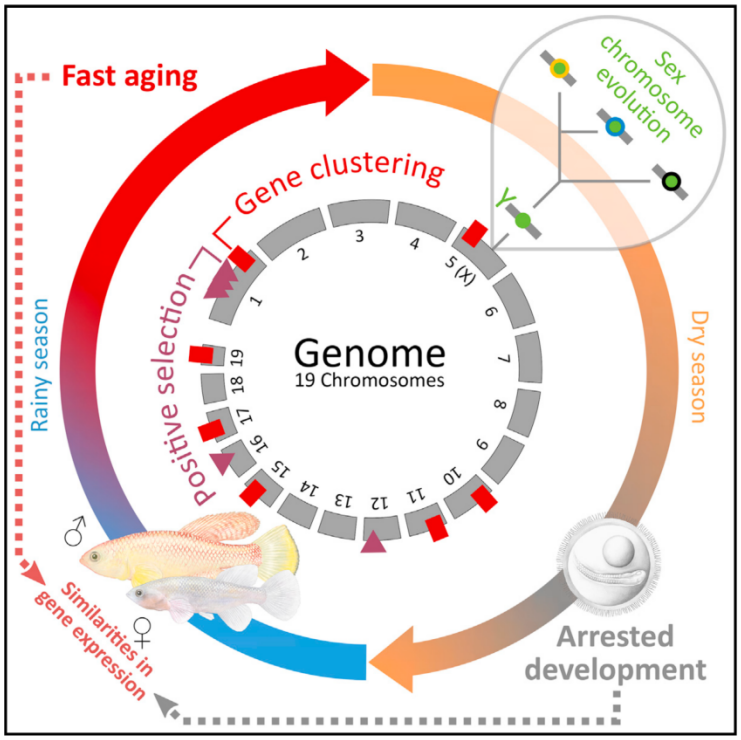
## Manuskript II: Insights into Sex Chromosome Evolution and Aging from the Genome of a Short-Lived Fish

Resource

Cell

# Insights into Sex Chromosome Evolution and Aging from the Genome of a Short-Lived Fish

## Graphical Abstract




## Highlights

- The genome sequence of a very short-lived fish is a resource for aging research
- The sex chromosomes display features of early mammalian XY evolution
- Aging-related genes are clustered in specific genomic regions
- Transcriptional profiles show similarities between developmental arrest and aging

## Authors

Kathrin Reichwald, Andreas Petzold, Philipp Koch, ..., Alessandro Cellerino, Christoph Englert, Matthias Platzer

## Correspondence

matthias.platzer@leibniz-fli.de

## In Brief

The turquoise killifish has a lifespan of only 4–12 months and yet its aging shares many similarities with that of humans. We sequenced and analyzed the killifish genome and provide insights into its biology. We detected very early stages of sex chromosome evolution, identified the sex-determining master gene, found clustering of aging-related genes in the genome, identified genes under positive selection, and discovered that similar gene sets are regulated during developmental arrest of embryos and aging.

## Accession Numbers

KG817100
KG959958

Reichwald et al., 2015, Cell *163*, 1527–1538
December 3, 2015 ©2015 Elsevier Inc.
http://dx.doi.org/10.1016/j.cell.2015.10.071

Resource

Cell

# Insights into Sex Chromosome Evolution and Aging from the Genome of a Short-Lived Fish

Kathrin Reichwald,[1,14] Andreas Petzold,[1,14,16] Philipp Koch,[1,14] Bryan R. Downie,[1,14] Nils Hartmann,[1,14] Stefan Pietsch,[1] Mario Baumgart,[1] Domitille Chalopin,[2,17] Marius Felder,[1] Martin Bens,[1] Arne Sahm,[1] Karol Szafranski,[1] Stefan Taudien,[1] Marco Groth,[1] Ivan Arisi,[3] Anja Weise,[4] Samarth S. Bhatt,[4] Virag Sharma,[5,6] Johann M. Kraus,[7] Florian Schmid,[7,8] Steffen Priebe,[9] Thomas Liehr,[4] Matthias Görlach,[1] Manuel E. Than,[1] Michael Hiller,[5,6] Hans A. Kestler,[1,7,10] Jean-Nicolas Volff,[2] Manfred Schartl,[11,12] Alessandro Cellerino,[1,13,15] Christoph Englert,[1,10,15] and Matthias Platzer[1,15,*]

[1]Leibniz Institute on Aging-Fritz Lipmann Institute (FLI), Jena 07745, Germany
[2]Institut de Génomique Fonctionnelle de Lyon, Ecole Normale Supérieure de Lyon, CNRS UMR5242, Université Claude Bernard Lyon 1, 69364 Lyon Cedex, France
[3]Genomics Facility, European Brain Research Institute (EBRI) Rita Levi-Montalcini, Rome 00143, Italy
[4]Jena University Hospital, Institute of Human Genetics, Friedrich Schiller University, Jena 07743, Germany
[5]Max Planck Institute of Molecular Cell Biology and Genetics, Dresden 01307, Germany
[6]Max Planck Institute for the Physics of Complex Systems, Dresden 01307, Germany
[7]Medical Systems Biology, Ulm University, Ulm 89069, Germany
[8]International Graduate School in Molecular Medicine at Ulm University (GSC270), Ulm 89069, Germany
[9]Leibniz Institute for Natural Product Research and Infection Biology-Hans-Knoell-Institute (HKI), Jena 07745, Germany
[10]Faculty of Biology and Pharmacy, Friedrich Schiller University Jena, Jena 07743, Germany
[11]Department of Physiological Chemistry, Biocenter, University of Wuerzburg, Wuerzburg 97074, Germany
[12]Comprehensive Cancer Center Mainfranken, University Hospital Wuerzburg, Wuerzburg 97074, Germany
[13]Laboratory of Biology, Scuola Normale Superiore, Pisa 56126, Italy
[14]Co-first author
[15]Co-senior author
[16]Present address: Deep Sequencing Group SFB 655, Biotechnology Center, Dresden University of Technology, Dresden 01307, Germany
[17]Present address: Department of Genetics, University of Georgia, Athens, GA 30602, USA
*Correspondence: matthias.platzer@leibniz-fli.de
http://dx.doi.org/10.1016/j.cell.2015.10.071

## SUMMARY

The killifish *Nothobranchius furzeri* is the shortest-lived vertebrate that can be bred in the laboratory. Its rapid growth, early sexual maturation, fast aging, and arrested embryonic development (diapause) make it an attractive model organism in biomedical research. Here, we report a draft sequence of its genome that allowed us to uncover an intra-species Y chromosome polymorphism representing—in real time—different stages of sex chromosome formation that display features of early mammalian XY evolution "in action." Our data suggest that *gdf6Y*, encoding a TGF-β family growth factor, is the master sex-determining gene in *N. furzeri*. Moreover, we observed genomic clustering of aging-related genes, identified genes under positive selection, and revealed significant similarities of gene expression profiles between diapause and aging, particularly for genes controlling cell cycle and translation. The annotated genome sequence is provided as an online resource (http://www.nothobranchius.info/NFINgb).

## INTRODUCTION

The turquoise killifish *Nothobranchius furzeri* (Jubb, 1971) is an annual fish that inhabits seasonal freshwater ponds in the south-east of Africa. It is characterized by rapid growth, early sexual maturation, and an exceptionally short lifespan reflecting the adaptation to the ephemeral nature of the habitat (Blažek et al., 2013; Cellerino et al., 2015; Genade et al., 2005). Several laboratory strains exist differing in their origin and lifespan. The GRZ strain comes from a semi-arid habitat in Zimbabwe (Figures 1A, 1C, and 2A) (Jubb, 1971), where its founders were collected in 1969 and have a maximum lifespan of 4–6 months. To date, this is the shortest maximum lifespan reported for a vertebrate bred in captivity (Valdesalici and Cellerino, 2003). Strains from semi-arid or more humid regions in Mozambique (e.g., MZM-0403 and MZM-0410) (Figure 1B) and the borderland between Mozambique and Zimbabwe (MZZW-0701) have a longer maximum lifespan of ~1 year (Terzibasi et al., 2008; Tozzini et al., 2013). These strains were established only several years ago and are genetically heterogeneous, whereas GRZ is highly inbred (Reichwald et al., 2009). In spite of the short lifespan, both GRZ and MZM strains show typical signs of aging, i.e., a decline in cognitive and behavioral capacity accompanied by aging-related histological changes (Di Cicco et al., 2011; Terzibasi et al., 2007) as well as aging-related telomere shortening and impairment of mitochondrial function (Hartmann et al.,

Cell

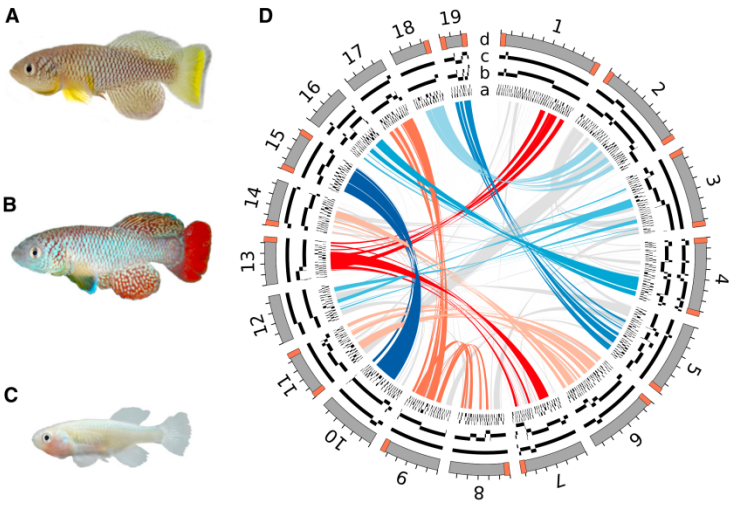



**Figure 1. The Turquoise Killifish and the Genome Assembly**

(A) Adult GRZ male.

(B) Adult MZM-0403 male.

(C) Adult GRZ female.

(D) Circles: the stepwise assembly of the reference sequence is represented from the inner to the outer circle. a: scaffolds obtained by applying programs ALLPATHS-LG and KILAPE. b: super-scaffolds built upon integration of optical mapping data. c: genetic scaffolds generated by linkage map integration. d: synteny groups defined upon analyses of synteny in medaka and stickleback. Synteny groups are sorted by length and numbered accordingly. Chromosome ends identified by optical mapping are marked in orange. The distance between two ticks is 10 Mb. Center: pairs of paralogous genes for synteny groups with a 1:1 (1:2) relation are connected by blue (red) lines; different hues define different chromosomal pairs (trios). Grey lines indicate gene pairs that do not follow our classification of chromosomal paralogy.

See also Figures S1 and S2 and Data S1.

2009, 2011). Lifespan determination in *N. furzeri* is polygenic; four quantitative loci relevant for lifespan are presently known (Kirschner et al., 2012).

Due to its fast development, *N. furzeri* can reach sexual maturity in <3 weeks and first signs of sexual dimorphism are apparent at 2 weeks after hatching (Blažek et al., 2013). In vertebrates, the gonads are usually the last organ system to develop into the functional adult structure. In fish, gonad differentiation commences only at late larval stages or even after metamorphosis and full functionality is reached at puberty (Devlin and Nagahama, 2002). Also, the sex determination system that provides the decision whether the undifferentiated gonad anlage of the embryo will develop later into a testis or an ovary is very plastic and can differ between closely related fish species or even within species (Volff et al., 2007). The fact that sex determination systems can change easily or arise rapidly during fish evolution together with the necessary rapid development of the reproductive system observed in *N. furzeri*, raises the question whether fast lifecycle and short lifespan influenced the evolution of the primary sex-determining (SD) gene and the sex chromosomes. Thus far, the segregation analyses of four sex-linked markers in crosses of GRZ and MZM-0403 is concordant with an XY sex determination system (Kirschner et al., 2012; Valenzano et al., 2009). The identical morphology (homomorphy) of the putative sex chromosomes (Reichwald et al., 2009) pointed to their young age and possibly to a situation of "sex chromosome evolution in action."

To survive the dry season, embryos of *N. furzeri* are protected from dehydration by a desiccation-resistant chorion and can enter into a state of developmental arrest termed diapause; the latter being a well-known adaptation in animal species to overcome unfavorable conditions. In *N. furzeri*, the arrest may occur at three distinct developmental stages (diapause I, II, and III) and can last for more than a year. Also in the nematode *Caenorhabditis elegans*, a larval arrest is observed (dauer larvae), and genes relevant for entering and maintaining the dauer state affect lifespan (Kenyon et al., 1993). We therefore analyzed whether differentially expressed genes (DEGs) in *N. furzeri* diapause versus non-diapause embryos are regulated in aging.

Recently, protocols for transgenesis (Hartmann and Englert, 2012; Valenzano et al., 2011) and CRISPR/Cas9-mediated mutagenesis have been established for *N. furzeri* (Harel et al., 2015). These tools, together with the short lifespan, make *N. furzeri* a very attractive vertebrate model to study aging, developmental arrest, and the interrelationship between both phenotypes. Here, we report a high-quality draft sequence of the *N. furzeri* genome. We provide insights into the very early evolutionary stages of an XY sex determination system, reveal clustering of aging-related genes in specific genomic regions, identify genes under positive selection and detect common expression profiles in diapause and aging.

## RESULTS AND DISCUSSION

### Assembly and Annotation of a High-Quality Draft Genome Sequence with Long-Range Contiguity

Today's challenge in genome analysis is generating a reference sequence of high quality and long-range contiguity. The *N. furzeri* project required special efforts because the genome is large and repeat-rich (Reichwald et al., 2009). In these two aspects, it resembles the zebrafish genome, for which a high-quality reference sequence was published only recently (Howe et al., 2013). We sequenced genomic DNA from *N. furzeri* females of the highly inbred GRZ strain (Figures 1A and 1C) in which all autosomes and the X chromosome are nearly homozygous. Using Illumina and Roche next-generation sequencing (NGS) technologies, we obtained whole-genome shotgun (WGS) data from 17 paired-end and mate-pair libraries amounting to 236 Gb (158-fold coverage, based on a genome-size estimate of 1.5 Gb; Figure S1B; Data S1A and S1B). Further, we sequenced

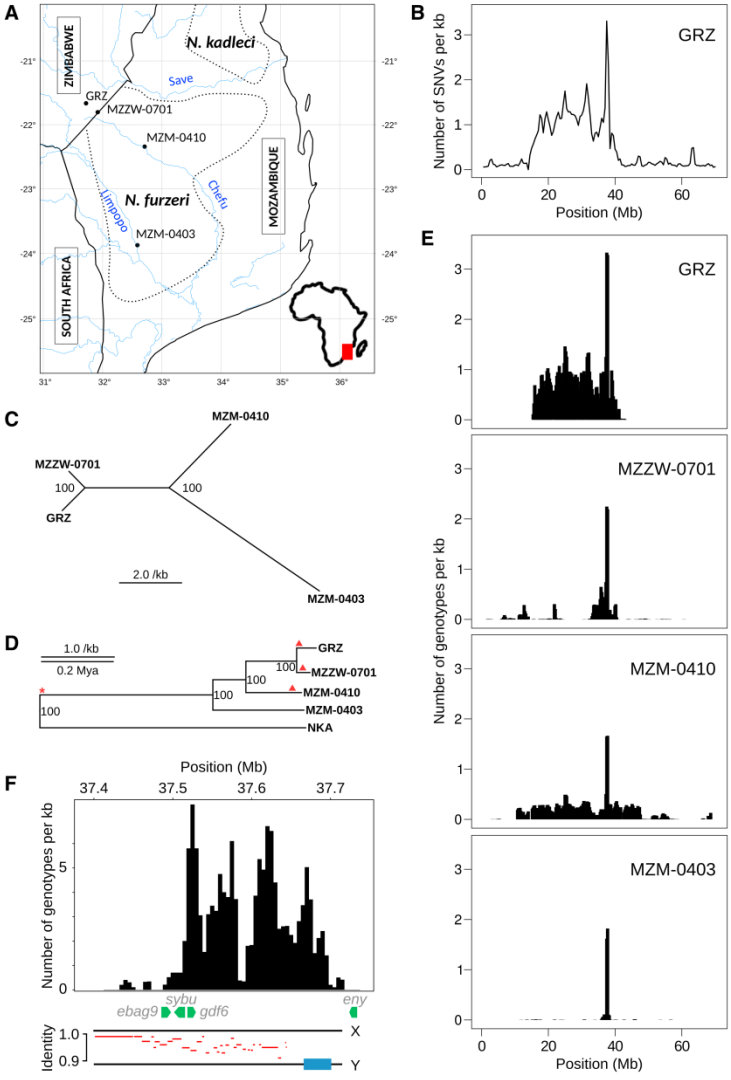



**Figure 2. Phylogeny and Sex Chromosome Analyses of *N. furzeri* Strains GRZ, MZZW-0701, MZM-0410, and MZM-0403**

(A) Geographic origin of *N. furzeri* strains is indicated by dots. The distribution range of *N. furzeri* and *N. kadleci* is marked by dotted lines.

(B) SNV density profile for synteny group (sgr) 05 obtained by aligning GRZ male WGS reads to the female reference sequence (sliding window: 1 Mb, step size: 500 kb).

(C) Phylogenetic relationship between strains based on WGS variation data.

(D) Phylogenetic tree of *N. furzeri* strains rooted by their sister species *N. kadleci* (NKA) based on exonic variations obtained by RNA-seq (Data S4I). The divergence time of *N. furzeri* and *kadleci* was estimated as 0.75 Mya (Dorn et al., 2014) and used for scaling. Red marks indicate the primary (asterisk) and secondary (triangle) events leading to the suppression of recombination shown in (E) and (F).

(E) Genotype density profiles for sgr05 of the strains. Genotype data were filtered for SNV positions in a given strain where all females are homozygous and all males are heterozygous (sliding window: 500 kb, step size: 250 kb).

(F) Top: zoom into the genotype density profile of the sex-determining region (SDR) in MZM-0403 (sliding window: 10 kb, step size: 5 kb). Genes annotated in the SDR of the GRZ female reference sequence are shown as green arrows. Bottom: identity plot (red lines) of BAC-based X and Y chromosome-specific sequences (black lines). The blue box represents a Y-specific 35 kb tandem repeat cluster composed of repeat units of 634 nt and 150 nt.

See also Figure S3 and Data S1I and S2A–S2C.

genomic insert ends of 81,393 BACs and fosmids to assist in the assembly and to provide a physical resource of the *N. furzeri* genome (5.3-fold clone coverage; Data S1C and S1D). To build the assembly, a five-step strategy was applied: we started with ALLPATHS-LG (Gnerre et al., 2011), continued with scaffolding, integrated optical and three genetic linkage maps, and finished with comparative synteny mapping in two closely related fish species (Table 1). The incorporation of optical mapping data remarkably improved the assembly contiguity (30-fold, Figure S1C). The genome assembly, in the following referred to as reference sequence, comprises 1.24 Gb (scaffold N50: ~0.5 Mb, optical N50: ~16 Mb, synteny N50: ~57 Mb), of which 87% were assigned to 19 synteny groups (sgrs) (Figure 1D). For 15 sgrs, we identified the corresponding *N. furzeri* chromosomes by fluorescence in situ hybridization using BAC probes (Figure S2A; Data S1E). Further, optical mapping data indicate that the assembly reached 22 of 38 chromosome ends (Figures 1D and S1D).

We built a comprehensive catalog of repetitive elements using Sanger/Illumina WGS reads and the genome assembly. Based on the Sanger data, we determined a repeat content of 64.6%, comprising 42.1% dispersed and 22.5% tandem repeats. This was confirmed by non-assembled NGS WGS data (Figure S2B). The *N. furzeri* reference sequence, however, contains only 35% repeats. In particular, tandem repeats are underrepresented (2% instead of 22.5%). This is most likely caused by the short NGS reads that collapse during the assembly process. Dispersed repeats amount to 33%, with LINEs being most abundant (8.4%) attributable to a recent expansion in this class of retrotransposons (Figure S2C; Data S1F). Finally, we confirmed the high quality of the reference sequence by PacBio WGS- and BAC sequencing (Data S1H and S1I) showing that gaps

Cell

**Table 1. Statistics of the Stepwise Assembly**

| Assembly Step | Number of Scaffolds | Total Length (bp) | Fraction of N[a] (%) | Longest Assembly Unit (bp) | N50 (bp) |
|---|---|---|---|---|---|
| A ALLPATHS-LG | 15,930 | 900,823,930 | 9.9 | 1,451,049 | 132,538 |
| B Scaffolding + gap filling | 7,675 | 943,595,854 | 9.2 | 3,869,209 | 494,454 |
| C Optical map integration | 6,012 | 1,230,898,532 | 30.4 | 44,272,285 | 15,858,201 |
| D Genetic map integration | 5,924 | 1,239,698,532 | 30.9 | 96,068,516 | 48,234,189 |
| E Synteny integration | 5,896 | 1,242,498,532 | 31.0 | 98,476,147 | 57,367,160 |
| Anchoring within the Final Assembly | | | | | |
| Chromosomes/synteny groups | 19 | 1,078,719,814 | 33.64 | 98,476,147 | 63,666,967 |
| Autosomes | 18 | 1,008,464,687 | 33.42 | 98,476,147 | 57,680,405 |
| X chromosome | 1 | 70,255,127 | 36.78 | 70,255,127 | 70,255,127 |
| Unassigned | 5,877 | 163,778,718 | 13.94 | 1,706,182 | 81,864 |

See also Data S1.
[a]Unresolved nucleotide positions, stands for A, C, G, or T.

contained in the genome assembly are almost entirely composed of repeats (83.1%).

We performed gene annotation using comprehensive RNA sequencing (RNA-seq) and microRNA sequencing (miRNA-seq) datasets as well as protein homology and in silico prediction tools (Figure S2D; Data S1K–S1N). We annotated 26,141 protein-coding genes with 59,154 transcripts, and 59 rRNA, 453 tRNA, 184 small nucleolar RNA (snoRNA), 598 miRNA, and 117 other non-protein coding RNA (ncRNA) genes (a detailed description of the miRNome will be reported elsewhere; M. Baumgart, I.A., and A.P., unpublished data). The teleost genome duplication (TGD) is reflected by the presence of 2,229 paralogous gene pairs, representing 17% of the *N. furzeri* protein-coding genes; further, we identified five pairs of putatively paralogous chromosomes with a 1:1 and three triads with a 1:2 relationship (Figure 1D; Data S1O).

To assess the completeness of the reference sequence with respect to the non-repetitive fraction of the genome, we used the Core Eukaryotic Genes Mapping Approach (CEGMA) (Parra et al., 2007) and searched in *N. furzeri* for orthologs of 248 highly conserved genes present in most eukaryotic genomes. Of these, we detected 98% in the reference sequence with 95% being completely covered (Data S1P). Furthermore, we could align 91% of the *N. furzeri* transcript catalog (Petzold et al., 2013) with the reference sequence strongly suggesting a highly complete representation of the genic fraction of the genome. Moreover, the PacBio-based estimate of the repeat content in gaps confirms that ~90% of the non-repetitive genome fraction is represented in the assembly. The annotated genome reference sequence is accessible at the *N. furzeri* Information Network Genome Browser (NFINgb, http://www.nothobranchius.info/NFINgb). Its long-range contiguity, chromosomal scale assembly, and completeness of genic regions allow studying the biology of the *N. furzeri* genome.

## Insights into Early Events of XY Sex Chromosome Evolution

To map the SD region (SDR) in the reference sequence, which represents a GRZ female genome, we performed additional WGS sequencing of four GRZ males (Data S2B). Because the GRZ strain is highly inbred, we expected genomic variations predominantly in the region of suppressed recombination between male and female sex chromosomes. Accordingly, male single nucleotide variations (SNVs) were mainly confined to a region on sgr05 (Figures 2B and S3A) that bears the only four sex-linked markers identified so far (Kirschner et al., 2012; Valenzano et al., 2009). This male-specific region of the Y chromosome (MSY) encompasses 26.1 Mb (sgr05: 15,031,832–41,162,746) and exhibits a distinct peak in variation density at position 37.6 Mb. PCR/Sanger sequencing-based validation of sex-linkage for selected SNVs pointed to an intra-species sex chromosome polymorphism between *N. furzeri* strains. For example, variations in the syntabulin gene (*sybu*) are associated with sex in GRZ, MZZW-0701, and MZM-0410 but not in MZM-0403, whereas SNVs up to 42 kb upstream of *sybu* show sex-linkage in all strains (Data S2A).

By analyzing the intra-species variations by additional WGS data from males and females of MZZW-0701, MZM-0410, and MZM-0403 in more detail (Data S2B), we identified ~3.3 million SNVs (accessible at NFINgb). Using those SNVs to determine the phylogenetic relationship between strains, we found a good agreement with the geographic location of collection sites (Figures 2A and 2C). Rooting of the phylogenetic tree revealed that MZM-0403 belongs to a different lineage than the three other strains (Figure 2D), thus confirming the deep geographic structuring of the species (Bartáková et al., 2013; Dorn et al., 2011). We next searched genome-wide for signs of suppressed recombination and identified the most prominent region in all strains on sgr05 (Figures 2E and S3A). In GRZ, the SNV and genotype density profiles coincide (Figures 2B and 2E) suggesting that the same genetic signal of suppressed sex chromosomal recombination was detected with both approaches. While the size of the MSY differs considerably between strains, ranging from 196 kb to 37 Mb (Data S2C), the position of the variation peak is identical. To date, intra-species sex chromosome polymorphisms have been observed only in exceptional cases and only by using cytogenetic methods, e.g., in guppy (Nanda et al., 2014).

Cell

Comparative variation analyses of this remarkable strain-specific Y chromosome polymorphism indicate a two-step scenario for its evolution. First, an ancient event in the common ancestor of all strains led to suppressed recombination in a 196-kb region and the emergence and/or fixation of a SD signal. This stage of early sex-chromosome evolution is conserved in MZM-0403 (Figure 2F, top). In all four strains, the highest number of sex chromosome-specific SNVs was accumulated in the 196-kb region, indicating that recombination suppression shielded in the ancestral state the newly evolving SD gene from cross-over and the proto-Y from losing its identity. To shed light on the mechanism of recombination suppression, we sequenced X- and Y-specific BACs harboring this region using PacBio technology (Data S1I). The BAC-based X-specific assemblies confirmed the reference sequence. In addition, we obtained a corresponding Y-specific region encompassing a 35 kb tandem-repeat cluster (Figure 2F, bottom) that similarly to the MSY of the medaka fish (Kondo et al., 2006) may prevent recombination in flanking regions.

Secondary events encompassing larger regions (7–37 Mb), yet containing the primary SDR, occurred independently in each of the three northern strains. By applying FISH analysis, we identified an inversion as the secondary cross-over barrier in MZM-0410 (Figure S3B). Thus, the individually structured *N. furzeri* Y chromosomes seem to reflect the first stages of the mammalian XY evolution that has shaped these chromosomes by consecutive inversions into evolutionary strata over 320 million years (Lahn and Page, 1999). Also, the sex chromosomes of the flatfish *Cynoglossus semilaevis* estimated to be ~30 million years old, have most likely diverged due to suppression recombination by a large inversion (Chen et al., 2014). For *N. furzeri* we estimate the occurrence of the secondary recombination suppression in GRZ around 70 thousand years ago (kya), in MZZW-0701 50 kya and in MZM-0410 38 kya by dating the primary event to the species split between *N. furzeri* and *N. kadleci* at 750 kya (Dorn et al., 2014) (Figure 2D; Data S2D). Although this is a rough estimate, we conclude that the secondary events are very young in evolutionary terms compared to previously studied SD systems.

Our data demonstrate that during early sex chromosome evolution, a whole set of different Ys can be created. In-depth analyses of Y polymorphisms in species with older Y chromosomes will allow studying whether in a second phase the most successful Y might make a sweep through the species. Such a sweep would then lead to a situation noticed for mammalian Ys where only minor sequence variations mark the Y haplotypes in a later phase of Y chromosome evolution (Ellegren, 2003). Future studies will clarify whether population-genetic fragmentation (Bartáková et al., 2013), short lifespan, annualism, and/or the multiple specific adaptations of *N. furzeri* facilitated its unprecedented Y chromosome polymorphism.

### Tracing the Emergence of a Novel Sex-Determining Gene: *gdf6Y*

We next attempted to identify the SD gene in *N. furzeri*. The minimal MSY was observed in MZM-0403 encompassing 196 kb and coinciding with the peak of Y-specific sequence variation at position 37.6 Mb in sgr05 (Figures 2E and 2F). This region contains only one annotated gene, *gdf6*, encoding growth differentiation factor 6, a member of the TGF-β family. We propose *gdf6Y* as symbol for the gene in the MSY. In GRZ, the *gdf6Y* coding sequence (CDS) differs from *gdf6* on the X chromosome in 22 SNVs and a 9-bp deletion, resulting in 15 amino acid (aa) exchanges and a 3 aa deletion (Figure S4A). All non-synonymous SNVs and the deletion are conserved between strains (Data S3A and S3B). Remarkably, the part of *gdf6Y* coding for the C-terminal 120 aa and homologous to the mature human GDF6, contains five non-synonymous but no synonymous substitutions indicating that positive selection acted on this part of the protein. The mature growth factor is highly conserved between vertebrates, and all male substitutions affect aa conserved between the *N. furzeri* X-chromosomal Gdf6 and its human ortholog (Figure 3A). Scanning all 339 genes in the 26.1 Mb MSY of GRZ confirms the sequence variations in *gdf6Y* as by far strongest signal of local positive selection (Data S3C).

To evaluate the impact of these aa changes, we performed homology modeling using the structure of the human receptor-bound GDF5 dimer (Kotzsch et al., 2009). Four of the five aa differing between mature Gdf6 and Gdf6Y reside in the modeled region (Figure 3A). Two of them (Gdf6Y/Gdf6: R405/Q408 and V407/E410) point outward into the solvent and reside at the edge of a β sheet (Figures 3B and S4B) that undergoes an induced fit upon formation of the GDF5:receptor complex (Kotzsch et al., 2009). The other two (T364/I367 and V372/M375) are located in a helix being part of the protomer interface but also contacting the receptor (Kotzsch et al., 2009). Hence, all four X/Y variable aa might have a bearing on protein interactions, either during dimerization or in the process of forming complexes with receptor(s).

Comparative analyses of *gdf6*/*gdf6Y* transcript levels revealed biallelic expression in early developmental stages of male and female GRZ and a significantly higher overall expression in males starting at day 3 post-hatching (Figures S4C and S4D; Data S3D). In RNA-seq data of adult ovaries, we found few *gdf6* reads, whereas in testes only *gdf6Y* mRNAs were detected at a considerable level. A possible explanation for the male-specific expression from the Y-chromosomal locus is a *gdf6Y*-specific deletion of 241 bp (sgr05: 37,526,406–37,526,646) in the 3′UTR including a potential mir-430 binding site (Figures S4E–S4G). In fish, mir-430 is an important regulator of germline-specific gene expression (Mishima et al., 2006). It is tempting to speculate that this deletion was the primary event marking the inception of the XY differentiation.

*Gdf6Y* expression peaks shortly after hatching; this is a time period when sex determination occurs in many fish species. Gdf6 is a member of the TGF-β family known to play a predominant role in developmental processes. Other members of the TGF-β family, e.g., the anti-Mullerian hormone (AMH) and the gonadal soma-derived growth factor (GSDF), as well as their receptors are important factors in sexual development of mammals and other vertebrates and function as master male sex determinants in several fish species (Josso and Clemente, 2003; Kikuchi and Hamaguchi, 2013; Morrish and Sinclair, 2002; Myosho et al., 2012; Rondeau et al., 2013). Gdf9 and Bmp15 are important players in ovarian development of mammals (Otsuka et al., 2011) and fish (Clelland and Kelly, 2011). Gdf6 has not

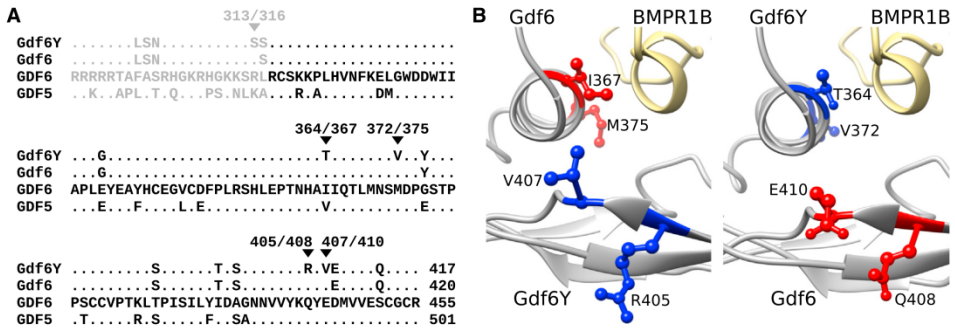



**Figure 3. Gdf6Y/Gdf6 Homology Modeling**

(A) ClustalW alignment of C-terminal, highly conserved 125 aa of *N. furzeri* Gdf6Y and Gdf6 as well as human GDF6 and GDF5. Amino acids (aa) identical to GDF6 are shown as dots. Amino acids varying between Gdf6Y and Gdf6 are highlighted by filled triangles and their numbers. The first 22 aa depicted in gray were not included in the modeling because they are missing in the reference structure.

(B) Detailed ribbon representations of two regions (left, right) of the modeled Gdf6Y/Gdf6 hetero-dimer (gray) receptor (yellow) complex given in Figure S4B. The four Gdf6Y/Gdf6 variable aa covered by the model are shown with side chains in blue for Gdf6Y and red for Gdf6. In the dimer, these aa are located spatially close to each other in the two regions shown.

See also Figure S4 and Data S3.

been described in the context of gonad development so far; how it acts as a master sex regulator in *N. furzeri* warrants further investigation.

## Genomic Positional Enrichment of Aging-Related Genes

Recently, data have accumulated suggesting that eukaryotic genes located in physical proximity may be co-regulated and/or have similar functions. Correlations between chromosomal position and membership of functional gene sets were identified for yeast (Santoni et al., 2013) and human (Thévenin et al., 2014) genomes. Hence, chromosomal and spatial co-localization in the nucleus may indicate co-regulation. It was previously shown that 3D chromatin structure couples nuclear compartmentalization of chromatin domains with the control of gene activity (Guelen et al., 2008) and thus contributes to cell-specific gene expression (Zullo et al., 2012). In this context, it is noteworthy that cellular senescence is associated with modifications of the global chromatin interaction network (Chandra et al., 2015). To our knowledge, it has not yet been investigated whether genes relevant for organismal aging are clustered in genomic regions.

Taking advantage of the long-range contiguity of the *N. furzeri* reference sequence, we set out to study whether aging-related genes show positional gene enrichment (PGE) in sgrs. To this end, we identified aging-related DEGs in three tissues (brain, liver, and skin) by applying two different approaches: (1) we compared young versus old MZM-0410 (5 weeks versus 39 weeks, corresponding to 10% versus 75% of maximum lifespan), and (2) we compared GRZ versus MZM-0410 at 12 weeks. As aging rates differ between these strains (Terzibasi et al., 2008), the same chronological age in the second approach corresponds to 50% of the maximum lifespan in GRZ and 24% in MZM-0410 (Data S4A–S4G).

In total, we detected ten PGE regions. Four of those are based on DEGs obtained by the first approach and six were identified by the second approach (false discovery rate [FDR] < 0.05, scan statistics; Data S4H). These regions are located on seven sgrs, extend over 2.6–9.2 Mb, and contain 11–23 DEGs. On three sgrs, two PGE regions each overlap non-randomly ($p = 0.0012$, resampling test) indicating that the same genomic features were detected by different approaches and in samples from different organs. One of the latter PGE regions located on sgr10 and detected based on DEGs in skin aging (Figure 4A) is enriched for the GO term "response to wounding" (FDR < 0.05, Fisher's exact test). The genes are downregulated in aging (Figure 4B) thus suggesting their co-regulation and providing a link to the well-accepted aging-related phenotype of decreased regenerative capacity (Conboy et al., 2005). These findings demonstrate that *N. furzeri* genes related to aging are distributed non-randomly in the genome and that positional clustering may allow their co-regulation.

## Positively Selected Genes in *N. furzeri*

The availability of high-quality genomic reference sequences facilitates the identification of genes under positive selection. To identify genes potentially relevant for adaptation of life-history traits we analyzed *N. furzeri* in comparison with *N. pienaari* because these sympatric species show convergent evolution of short lifespan (Tozzini et al., 2013). Therefore, we generated CDS data for *N. pienaari* and, additionally, for four longer-lived Nothobranchius species as well as the non-annual killifish *Aphyosemion striatum* as outgroup by RNA-seq of brain samples (Data S4I). The consensus tree based on multi-species CDS alignments matched well their reported phylogeny (Dorn et al., 2014) (Figure 4C). To avoid assembly errors, only de novo assembled *N. furzeri* transcripts that show 100% identity to the reference sequence (n = 23,108; corresponding to 11,748 genes) were analyzed. Accordingly, for *N. pienaari* we included transcripts showing at least 99% coverage and 98% identity to the *N. furzeri* reference sequence (n = 5,576; corresponding to 5,363 genes). We identified seven genes under

Cell

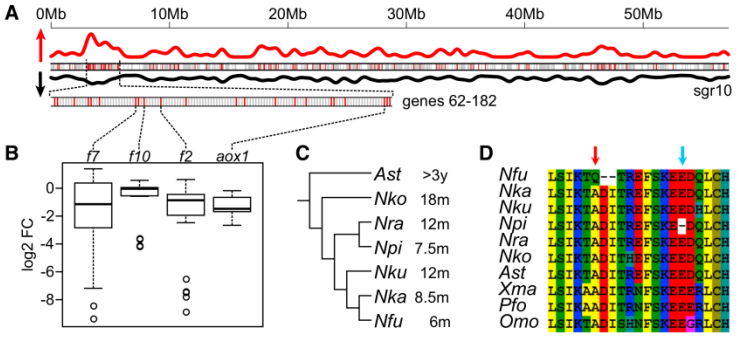



**Figure 4. Positional Gene Enrichment and Positive Selection**

(A) Schematic representation of synteny group (sgr) 10 and a region of positional gene enrichment. The genes in the sgr are represented by vertical bars: red, differentially expressed; gray, not differentially expressed. The density of all genes on the sgr (black line) and those differentially expressed (red line) is shown (kernel density estimation, Gaussian kernel). Arrows indicate the direction of increasing values.

(B) Relative downregulation of four DEGs with the GO annotation "response to wounding" in aging skin. Gene symbols *f2*, *f7*, and *f10* stand for coagulation factors II, VII, and X. *aox1*, aldehyde oxidase 1. Boxes, first and third quartiles; horizontal line, median; whiskers, most extreme value within 1.5× of inter-quartile range; dots, outliers. Expression differences were calculated by pairwise comparisons (n = 25) between the samples.

(C) Phylogram of the species used for transcriptome sequencing based on Dorn et al. (2014). For each species, the captive median lifespan is reported: *A. striatum* (unpublished), *N. korthause* (Baumgart et al., 2015), *N. rachovii*, *N. pienaari*, *N. kuhntae*, *N. furzeri* (Tozzini et al., 2013), and *N. kadleci* (Ng'oma et al., 2014).

(D) Alignment of the Id3 C terminus. The red arrow indicates aa under positive selection in *N. furzeri* followed by a two aa deletion. The blue arrow indicates the *N. pienaari*-specific deletion. The background color of each aa relates to the chemical nature of its side chain.

See also Data S4A–S4M.

positive selection in *N. furzeri* and one in *N. pienaari* (FDR < 0.05, Data S4J) highlighting the importance of a reference sequence for evolutionary analyses. Remarkably, five of these genes are either up- or downregulated in aging in at least one of three MZM-0410 organs (brain, liver, skin at 39 versus 5 weeks; Data S4A and S4K–S4M).

The signature of selection for *id3* (inhibitor of DNA binding 3, dominant negative helix-loop-helix protein) is particularly interesting. *Id3* is upregulated during aging in brain and skin and is also a key component of TGF-β signaling. TGF-β regulates inflammation, is involved in aging-related diseases such as tumorigenesis, fibrosis, glaucoma, and osteoarthritis (Krieglstein et al., 2012), and regulates life-history traits in *C. elegans* (Luo et al., 2010; Shaw et al., 2007). In *N. furzeri*, the gene shows signs of positive selection; i.e., a radical substitution of a non-polar by a charged aa followed by a 2-aa deletion (Figure 4D). Interestingly, at 10-aa distance in *N. pienaari* one evolutionarily conserved aa is deleted suggesting convergent evolution.

Another interesting gene under positive selection is *ikbip* (I Kappa B Kinase Interacting Protein), a pro-apoptotic gene (Hofer-Warbinek et al., 2004) downregulated in skin aging. Apoptosis is relevant for both diapause and aging. Diapausing killifish embryos are resistant to apoptosis (Meller and Podrabsky, 2013), but apoptosis is induced in aging *N. furzeri* (Di Cicco et al., 2011; Ng'oma et al., 2014). Apoptosis-related genes were shown to be age-regulated across tissues in a meta-analysis of mammalian aging (de Magalhães et al., 2009). Studies of larger taxonomical samples, including genomic and transcriptomic sequence datasets, are needed for further investigation of positive selection and convergent evolution in Nothobranchius species.

## Overlap of Transcriptional Changes in Developmental Arrest and Aging

Last, we assessed the potential relation between developmental arrest (diapause) and aging in *N. furzeri*. Focusing on diapause II at the somite stage, we determined gene expression changes between arrested and non-arrested embryos at a comparable morphological stage using RNA-seq (Figure 5A). We identified 1,256 down- and 971 upregulated genes in arrested GRZ and MZM-0403 embryos (FDR < 0.05, DEseq and edgeR; Data S4O). In the set of downregulated genes, pattern specification processes including embryonic development of different organs and processes associated with cell proliferation were enriched (p < 0.05, hypergeometric test). Processes enriched in upregulated genes were more diverse and included translational elongation, ribosome biogenesis, metabolism and regulation of cellular component movement (Figure S5A). Decreased rates of cell proliferation and changes in the metabolic status have also been observed in diapause embryos of the South American killifish *Austrofundulus limnaeus* (Podrabsky and Culpepper, 2012). Upregulation of genes involved in translational and ribosomal processes, however, was unexpected. A possible explanation is the need for immediate cellular activity once environmental conditions trigger the exit from diapause.

We then analyzed whether there were similar gene expression changes in diapause and aging. To this end, we again employed MZM-0410 RNA-seq data (brain, liver, skin; 5/12/20/27/39 weeks; Data S4A) and focused on genes showing a monotonic increase or decrease of transcript levels in aging (Data S4P–S4R). In brain, the number of genes that were either down- or upregulated in both aging and diapause was significantly higher than the number of genes downregulated in brain aging and upregulated in diapause or vice versa (p < 0.001, chi-square test; Figure 5B). We therefore concentrated on the first two groups with highest DEG numbers and found that all significantly enriched processes in the group of downregulated genes were associated with cell-cycle progression and DNA replication (Figure 5C). Previous work suggests that brain aging in *N. furzeri* is associated with reduced mitotic activity of adult neuronal stem cells (Tozzini et al., 2012). Unexpectedly, the

Cell

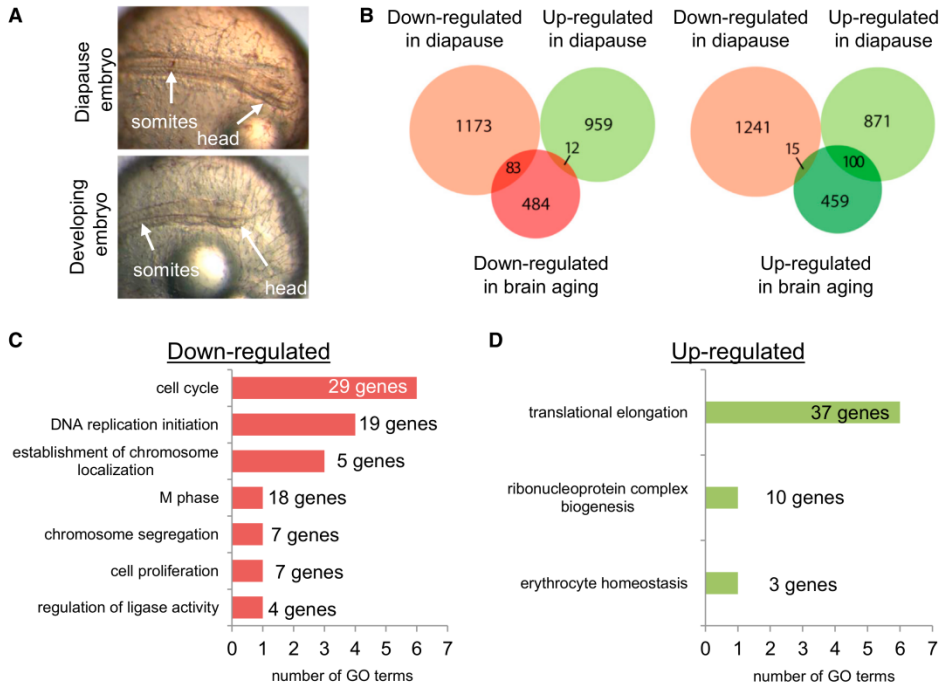



**Figure 5. RNA-Seq Analyses of Diapause Embryos and Brain Aging**

(A) The embryo (upper picture) has arrested in diapause II for 9 months, whereas the non-arrested embryo (lower picture) exhibiting a comparable morphological stage has an age of 6 days post fertilization.

(B) Venn-analyses of genes downregulated (light red) and upregulated (light green) in diapause as well as monotonic downregulated (dark red) and upregulated (dark green) in brain aging.

(C) Enrichment analyses of genes downregulated in diapause and brain aging. Numbers of involved genes and GO terms are shown for each biological process.

(D) Enrichment analyses of genes upregulated in diapause and brain aging.

See also Figure S5 and Data S4A and S4N–S4U.

two major processes enriched in upregulated genes in diapause and brain aging were translational elongation and ribonucleoprotein complex biogenesis (Figure 5D). The small number of overlapping DEGs between diapause and liver aging prevented further analysis (Figure S5B). Similar to brain, we identified in skin a significantly higher number of genes that were either up- or downregulated both in diapause and aging than genes regulated in opposite ways ($p < 0.001$, chi-square test, Figure S5B). Analysis of consistently downregulated genes showed enrichment of diverse processes. In the respective set of upregulated genes, however, again translational elongation and ribosome biogenesis were enriched (Figure S5C). Previously, aging-related upregulation of genes encoding translational and ribosomal proteins has been reported for human brain, muscle, and kidney suggesting a compensatory mechanism for aging-related increase in protein damage (Zahn et al., 2006). To our knowledge, a common expression profile for vertebrate developmental arrest and aging has not been described before.

In the nematode *C. elegans*, a link between developmental arrest, the so-called dauer larvae, and longevity has been identified. When mutated, some genes affecting dauer formation such as *daf-2* (a homolog of the insulin and IGF-1 receptor) increase lifespan (Kenyon et al., 1993; Lin et al., 1997; Ogg et al., 1997; Shaw et al., 2007). Moreover, the gene expression profile of dauer larvae shows similarities to the expression profile of long-lived adult mutants (McElwee et al., 2004). At present, the absence of long-lived mutants prevents such kinds of analysis in *N. furzeri*. Our comparison of gene expression changes between *N. furzeri* diapause embryos and *C. elegans* dauer larvae (Wang and Kim, 2003) revealed little overlap (Data S4V). This does not seem surprising, given the long evolutionary distance between the two species and their different habitats. The identification of e.g., *daf-16*/*FoxO4* being upregulated in embryonic arrest of both species, however, indicates commonalities between the two processes and calls for further analyses, e.g., genomic manipulation of the *FoxO4* locus in *N. furzeri*.

In conclusion, the high-quality draft sequence of the *N. furzeri* genome provided here and the availability of several *N. furzeri* strains that differ in lifespan represent excellent resources for studying and identifying genes involved in aging and longevity.

Cell

Furthermore, the novel genomic engineering tools now available in *N. furzeri* such as the CRISPR/Cas system (Harel et al., 2015) will allow the generation of mutant lines at a large scale providing a platform for drug screening and sophisticated models to study aging as well as aging-related and other diseases and to develop novel therapies.

## EXPERIMENTAL PROCEDURES

Additional details are provided in the Supplemental Experimental Procedures.

### Animal Material

Sample acquisition was carried out in accordance with the "principles of laboratory animal care" and the current version of the German Law on the Protection of Animals.

### De Novo Genome Sequencing and Assembly

Two adult female GRZ were sequenced using Illumina technology and assembled with ALLPATHS-LG. In parallel, two adult male GRZ were sequenced using Roche technology; these data served for long-range scaffolding and gap filling. Further, optical mapping (OpGen; http://www.opgen.com) was performed in one adult female GRZ. By combining restriction maps obtained with this procedure and sequence scaffolds, superscaffolds were formed. These were manually ordered in genetic scaffolds based on own genetic maps (Kirschner et al., 2012; Ng'oma et al., 2014). Finally, by synteny analyses in medaka and stickleback, genetic scaffolds were arranged in sgrs.

### Repeat Annotation

Repeats are identified by (1) RepeatModeler in the reference sequence, (2) RepeatMasker, RepeatScout (Price et al., 2005) for assembled Sanger sequences generated by whole-genome sample sequencing, and (3) RepARK (Koch et al., 2014) for WGS Illumina reads. Subsequently, libraries were merged in a *N. furzeri*-specific repeat library and finally used to annotate the reference sequence by RepeatMasker and TandemRepeatFinder (Benson, 1999).

### Gene Annotation and Identification of Paralogs

Protein-coding genes were annotated based on (1) ab initio gene prediction, (2) protein sequence similarity, and (3) Illumina RNA-seq data. Results were combined into CDS models with EVM and UTRs, and transcripts were constructed with PASA (Haas et al., 2008). Gene symbols and functions were annotated using homologous proteins of medaka, platyfish, stickleback, tetraodon, and zebrafish obtained from Ensembl (Cunningham et al., 2015). InterProScan75 (Zdobnov and Apweiler, 2001) was used to identify protein domains and to retrieve Gene Ontology annotations.

MiRNA genes were identified from Illumina miRNA-seq data. To detect rRNA genes, BLAT searches using known *N. furzeri* rRNA sequences (Reichwald et al., 2009) as queries were performed. In addition, miRNA, tRNA, rRNA, and other non-protein-coding genes were identified using ab initio gene prediction tools.

TGD-derived paralogs were identified with Ensembl Compara. First, *N. furzeri* genes were used to find orthologs in medaka, platyfish, stickleback, tetraodon, and zebrafish. Next, Ensembl gene IDs served as queries in Ensembl Compara to detect pairwise paralogous relationships. Any pair of duplicated genes originating before the teleost split was discarded. Finally, *N. furzeri* genes related to the same orthologous gene were also included.

### Genomic Resequencing of *N. furzeri* Strains and Variation Calling

Illumina WGS reads generated for all strains were mapped to the reference sequence with Bowtie2 (Langmead and Salzberg, 2012) (minimum mapping quality score of 11). Regions with alignment gaps were realigned with GATK (McKenna et al., 2010) and duplicate reads marked with Picard Tools (http://picard.sourceforge.net). Sequence variations and genotypes were called with GATK. Selected genomic regions were resequenced in additional specimens by PCR and Sanger technology as described (Reichwald et al., 2009).

### Overrepresentation Analysis

Zebrafish orthologs of *N. furzeri* genes were retrieved using BLAST. Human orthologs were fetched with R package orthology. GO enrichment analysis was done using DAVID (Huang et al., 2009) and summarized by REViGO (Supek et al., 2011).

### Positional Gene Enrichment

Aging-related DEGs were identified by Illumina RNA-seq. Scan statistics (Glaz et al., 2001) were used to test if an observed accumulation of k DEGs on a sgr containing N genes is likely to happen by chance. The scan statistic S is the maximal k in any interval W of fixed size w (w = 0.1 × N). Subsequently, an overrepresentation analysis for each detected genomic region was performed.

### Positive Selection

Protein-coding sequences of *N. kadleci*, *N. korthausae*, *N. kuhntae*, *N. pienaari*, *N. rachovii*, *A. striatum*, and *N. furzeri* were assembled de novo using Illumina RNA-seq data. Prank (Löytynoja and Goldman, 2008) alignments of orthologous CDS were filtered by Gblocks (Talavera and Castresana, 2007) and in-house software. Then, the improved branch-site test of positive selection was applied as described (Zhang et al., 2005). Ka/Ks ratios were calculated for all CDS pairs in the SDR both in total and in 333 nt windows sampled using a step size of 99 nt.

### Gene Expression Analysis in Diapause Embryos

In total, 287 diapause and 239 non-diapause embryos were collected at the somite stage. Approximately 30 embryos per state were pooled resulting in eight diapause and eight non-diapause samples. Total RNA was extracted and sequenced by Illumina RNA-seq. Significant DEGs were identified and an overrepresentation analysis was performed.

## ACCESSION NUMBERS

The accession number for the *N. furzeri* genome project including genome assembly and NGS data (WGS, RNA-seq, and BAC-seq) reported in this paper is BioProject: PRJEB5837. The accession numbers for the *N. furzeri* GRZ genomic insert end sequences of BACs and fosmids are GenBank: KG817100 to KG959958. The accession number for assembled Sanger WGS sequences is BioProject: PRJNA29535. Accession numbers of individual datasets are given in Data S1, S2, S3, and S4.

## SUPPLEMENTAL INFORMATION

Supplemental Information includes Supplemental Experimental Procedures, five figures, and four data sets and can be found with this article online at http://dx.doi.org/10.1016/j.cell.2015.10.071.

## AUTHOR CONTRIBUTIONS

C.E., K.R., and M.P. initiated, managed, and drove the genome project. K.R., N.H., M.Ba., S.T., and M.Gr. prepared the samples. K.R., S.T., and M.Gr. performed the sequencing. A.P., P.K., B.R.D., V.S., and M.H. performed the genome assembly and annotation. P.K., B.R.D., D.C., and J.N.V. performed the repeat analysis. M.Ba., M.Gr., A.C., and M.P. performed the mRNA analysis. I.A., M.Ba., A.P., and A.C. performed the miRNA analysis. K.R., A.W., S.S.B., and T.L. performed the chromosome FISH. K.R., A.P., P.K., M.F., K.S., N.H., M.S., C.E., and M.P. performed the sex chromosome evolution analysis. M.Go. and M.E.T. performed protein structure modeling. J.M.K., F.S., S.Pr., P.K., H.A.K., A.C., and M.P. performed the PGE analysis. A.S., M.Be., A.P., B.R.D., A.C., and M.P. performed the positive selection analysis. N.H., S.Pi., and C.E. performed the diapause analysis. All authors contributed to data interpretation. K.R., A.P., P.K., N.H., M.S., A.C., C.E., and M.P. wrote the manuscript.

## ACKNOWLEDGMENTS

We thank Silke Foerste, Ivonne Goerlich, Ivonne Heinze, Christin Hahn, Cornelia Luge, Sabine Matz, Martin Neumann, and Bernd Senf for technical

assistance. We thank Karl Lenhard Rudolph for discussions and Cornelia Platzer for critical reading of the manuscript. This work was supported by the Leibniz Association (WGL: PAKT-2006-FLI to C.E. and M.P., and SAW-2012-FLI to M.P.), the German Research Foundation (DFG: RE 3505/1-1 to K.R., HA 6214/2-1 to N.H., and SFB 1074 project Z1 to H.A.K.), the German Federal Ministry of Education and Research (BMBF: JenAge 0315581A/C to A.C., C.E., and M.P.; 031A099 to M.H.; Gerontosys II, Forschungskern SySTaR, project ID 0315894A to H.A.K.), the European Community's Seventh Framework Program (FP7/2007-2013 under grant agreement 602783 to H.A.K.), and the Italian Ministry of Higher Education (FIRB: RBAP10L8TY to I.A.).

Received: June 3, 2015
Revised: August 11, 2015
Accepted: October 21, 2015
Published: December 3, 2015

## REFERENCES

Bartáková, V., Reichard, M., Janko, K., Polačik, M., Blažek, R., Reichwald, K., Cellerino, A., and Bryja, J. (2013). Strong population genetic structuring in an annual fish, Nothobranchius furzeri, suggests multiple savannah refugia in southern Mozambique. BMC Evol. Biol. *13*, 196.

Baumgart, M., Di Cicco, E., Rossi, G., Cellerino, A., and Tozzini, E.T. (2015). Comparison of captive lifespan, age-associated liver neoplasias and age-dependent gene expression between two annual fish species: Nothobranchius furzeri and Nothobranchius korthause. Biogerontology *16*, 63–69.

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. *27*, 573–580.

Blažek, R., Polačik, M., and Reichard, M. (2013). Rapid growth, early maturation and short generation time in African annual fishes. Evodevo *4*, 24.

Cellerino, A., Valenzano, D.R., and Reichard, M. (2015). From the bush to the bench: the annual Nothobranchius fishes as a new model system in biology. Biol. Rev. Camb. Philos. Soc. Published online April 28, 2015. http://dx.doi.org/10.1111/brv.12183.

Chandra, T., Ewels, P.A., Schoenfelder, S., Furlan-Magaril, M., Wingett, S.W., Kirschner, K., Thuret, J.Y., Andrews, S., Fraser, P., and Reik, W. (2015). Global reorganization of the nuclear landscape in senescent cells. Cell Rep. *10*, 471–483.

Chen, S., Zhang, G., Shao, C., Huang, Q., Liu, G., Zhang, P., Song, W., An, N., Chalopin, D., Volff, J.N., et al. (2014). Whole-genome sequence of a flatfish provides insights into ZW sex chromosome evolution and adaptation to a benthic lifestyle. Nat. Genet. *46*, 253–260.

Clelland, E.S., and Kelly, S.P. (2011). Exogenous GDF9 but not Activin A, BMP15 or TGFβ alters tight junction protein transcript abundance in zebrafish ovarian follicles. Gen. Comp. Endocrinol. *171*, 211–217.

Conboy, I.M., Conboy, M.J., Wagers, A.J., Girma, E.R., Weissman, I.L., and Rando, T.A. (2005). Rejuvenation of aged progenitor cells by exposure to a young systemic environment. Nature *433*, 760–764.

Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., et al. (2015). Ensembl 2015. Nucleic Acids Res. *43*, D662–D669.

de Magalhães, J.P., Curado, J., and Church, G.M. (2009). Meta-analysis of age-related gene expression profiles identifies common signatures of aging. Bioinformatics *25*, 875–881.

Devlin, R.H., and Nagahama, Y. (2002). Sex determination and sex differentiation in fish: an overview of genetic, physiological, and environmental influences. Aquaculture *208*, 191–364.

Di Cicco, E., Tozzini, E.T., Rossi, G., and Cellerino, A. (2011). The short-lived annual fish Nothobranchius furzeri shows a typical teleost aging process reinforced by high incidence of age-dependent neoplasias. Exp. Gerontol. *46*, 249–256.

Dorn, A., Ng'oma, E., Janko, K., Reichwald, K., Polačik, M., Platzer, M., Cellerino, A., and Reichard, M. (2011). Phylogeny, genetic variability and colour polymorphism of an emerging animal model: the short-lived annual Nothobranchius fishes from southern Mozambique. Mol. Phylogenet. Evol. *61*, 739–749.

Dorn, A., Musilová, Z., Platzer, M., Reichwald, K., and Cellerino, A. (2014). The strange case of East African annual fishes: aridification correlates with diversification for a savannah aquatic group? BMC Evol. Biol. *14*, 210.

Ellegren, H. (2003). Levels of polymorphism on the sex-limited chromosome: a clue to Y from W? BioEssays *25*, 163–167.

Genade, T., Benedetti, M., Terzibasi, E., Roncaglia, P., Valenzano, D.R., Cattaneo, A., and Cellerino, A. (2005). Annual fishes of the genus Nothobranchius as a model system for aging research. Aging Cell *4*, 223–233.

Glaz, J., Naus, J., and Wallenstein, S. (2001). Scan Statistics (New York: Springer).

Gnerre, S., Maccallum, I., Przybylski, D., Ribeiro, F.J., Burton, J.N., Walker, B.J., Sharpe, T., Hall, G., Shea, T.P., Sykes, S., et al. (2011). High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc. Natl. Acad. Sci. USA *108*, 1513–1518.

Guelen, L., Pagie, L., Brasset, E., Meuleman, W., Faza, M.B., Talhout, W., Eussen, B.H., de Klein, A., Wessels, L., de Laat, W., and van Steensel, B. (2008). Domain organization of human chromosomes revealed by mapping of nuclear lamina interactions. Nature *453*, 948–951.

Haas, B.J., Salzberg, S.L., Zhu, W., Pertea, M., Allen, J.E., Orvis, J., White, O., Buell, C.R., and Wortman, J.R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. Genome Biol. *9*, R7.

Harel, I., Benayoun, B.A., Machado, B., Singh, P.P., Hu, C.K., Pech, M.F., Valenzano, D.R., Zhang, E., Sharp, S.C., Artandi, S.E., and Brunet, A. (2015). A platform for rapid exploration of aging and diseases in a naturally short-lived vertebrate. Cell *160*, 1013–1026.

Hartmann, N., and Englert, C. (2012). A microinjection protocol for the generation of transgenic killifish (Species: Nothobranchius furzeri). Dev. Dyn. *241*, 1133–1141.

Hartmann, N., Reichwald, K., Lechel, A., Graf, M., Kirschner, J., Dorn, A., Terzibasi, E., Wellner, J., Platzer, M., Rudolph, K.L., et al. (2009). Telomeres shorten while Tert expression increases during ageing of the short-lived fish Nothobranchius furzeri. Mech. Ageing Dev. *130*, 290–296.

Hartmann, N., Reichwald, K., Wittig, I., Dröse, S., Schmeisser, S., Lück, C., Hahn, C., Graf, M., Gausmann, U., Terzibasi, E., et al. (2011). Mitochondrial DNA copy number and function decrease with age in the short-lived fish Nothobranchius furzeri. Aging Cell *10*, 824–831.

Hofer-Warbinek, R., Schmid, J.A., Mayer, H., Winsauer, G., Orel, L., Mueller, B., Wiesner, Ch., Binder, B.R., and de Martin, R. (2004). A highly conserved proapoptotic gene, IKIP, located next to the APAF1 gene locus, is regulated by p53. Cell Death Differ. *11*, 1317–1325.

Howe, K., Clark, M.D., Torroja, C.F., Torrance, J., Berthelot, C., Muffato, M., Collins, J.E., Humphray, S., McLaren, K., Matthews, L., et al. (2013). The zebrafish reference genome sequence and its relationship to the human genome. Nature *496*, 498–503.

Huang, W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat. Protoc. *4*, 44–57.

Josso, N., and Clemente, Nd. (2003). Transduction pathway of anti-Müllerian hormone, a sex-specific member of the TGF-beta family. Trends Endocrinol. Metab. *14*, 91–97.

Jubb, R.A. (1971). A new Nothobranchius (Pisces, Cyprinodontidae) from Southeastern Rhodesia. J Am Killifish Association *8*, 12–19.

Kenyon, C., Chang, J., Gensch, E., Rudner, A., and Tabtiang, R. (1993). A C. elegans mutant that lives twice as long as wild type. Nature *366*, 461–464.

Kikuchi, K., and Hamaguchi, S. (2013). Novel sex-determining genes in fish and sex chromosome evolution. Dev. Dyn. *242*, 339–353.

Kirschner, J., Weber, D., Neuschl, C., Franke, A., Böttger, M., Zielke, L., Powalsky, E., Groth, M., Shagin, D., Petzold, A., et al. (2012). Mapping of quantitative trait loci controlling lifespan in the short-lived fish Nothobranchius furzeri–a new vertebrate model for age research. Aging Cell *11*, 252–261.

Koch, P., Platzer, M., and Downie, B.R. (2014). RepARK–de novo creation of repeat libraries from whole-genome NGS reads. Nucleic Acids Res. *42*, e80.

Kondo, M., Hornung, U., Nanda, I., Imai, S., Sasaki, T., Shimizu, A., Asakawa, S., Hori, H., Schmid, M., Shimizu, N., and Schartl, M. (2006). Genomic organization of the sex-determining and adjacent regions of the sex chromosomes of medaka. Genome Res. *16*, 815–826.

Kotzsch, A., Nickel, J., Seher, A., Sebald, W., and Müller, T.D. (2009). Crystal structure analysis reveals a spring-loaded latch as molecular mechanism for GDF-5-type I receptor specificity. EMBO J. *28*, 937–947.

Krieglstein, K., Miyazono, K., ten Dijke, P., and Unsicker, K. (2012). TGF-β in aging and disease. Cell Tissue Res. *347*, 5–9.

Lahn, B.T., and Page, D.C. (1999). Four evolutionary strata on the human X chromosome. Science *286*, 964–967.

Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat. Methods *9*, 357–359.

Lin, K., Dorman, J.B., Rodan, A., and Kenyon, C. (1997). daf-16: An HNF-3/forkhead family member that can function to double the life-span of Caenorhabditis elegans. Science *278*, 1319–1322.

Löytynoja, A., and Goldman, N. (2008). Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science *320*, 1632–1635.

Luo, S., Kleemann, G.A., Ashraf, J.M., Shaw, W.M., and Murphy, C.T. (2010). TGF-β and insulin signaling regulate reproductive aging via oocyte and germline quality maintenance. Cell *143*, 299–312.

McElwee, J.J., Schuster, E., Blanc, E., Thomas, J.H., and Gems, D. (2004). Shared transcriptional signature in Caenorhabditis elegans Dauer larvae and long-lived daf-2 mutants implicates detoxification system in longevity assurance. J. Biol. Chem. *279*, 44533–44543.

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303.

Meller, C.L., and Podrabsky, J.E. (2013). Avoidance of apoptosis in embryonic cells of the annual killifish Austrofundulus limnaeus exposed to anoxia. PLoS ONE *8*, e75837.

Mishima, Y., Giraldez, A.J., Takeda, Y., Fujiwara, T., Sakamoto, H., Schier, A.F., and Inoue, K. (2006). Differential regulation of germline mRNAs in soma and germ cells by zebrafish miR-430. Curr. Biol. *16*, 2135–2142.

Morrish, B.C., and Sinclair, A.H. (2002). Vertebrate sex determination: many means to an end. Reproduction *124*, 447–457.

Myosho, T., Otake, H., Masuyama, H., Matsuda, M., Kuroki, Y., Fujiyama, A., Naruse, K., Hamaguchi, S., and Sakaizumi, M. (2012). Tracing the emergence of a novel sex-determining gene in medaka, Oryzias luzonensis. Genetics *191*, 163–170.

Nanda, I., Schories, S., Tripathi, N., Dreyer, C., Haaf, T., Schmid, M., and Schartl, M. (2014). Sex chromosome polymorphism in guppies. Chromosoma *123*, 373–383.

Ng'oma, E., Reichwald, K., Dorn, A., Wittig, M., Balschun, T., Franke, A., Platzer, M., and Cellerino, A. (2014). The age related markers lipofuscin and apoptosis show different genetic architecture by QTL mapping in short-lived Nothobranchius fish. Aging (Albany, N.Y.) *6*, 468–480.

Ogg, S., Paradis, S., Gottlieb, S., Patterson, G.I., Lee, L., Tissenbaum, H.A., and Ruvkun, G. (1997). The Fork head transcription factor DAF-16 transduces insulin-like metabolic and longevity signals in C. elegans. Nature *389*, 994–999.

Otsuka, F., McTavish, K.J., and Shimasaki, S. (2011). Integral role of GDF-9 and BMP-15 in ovarian function. Mol. Reprod. Dev. *78*, 9–21.

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics *23*, 1061–1067.

Petzold, A., Reichwald, K., Groth, M., Taudien, S., Hartmann, N., Priebe, S., Shagin, D., Englert, C., and Platzer, M. (2013). The transcript catalogue of the short-lived fish Nothobranchius furzeri provides insights into age-dependent changes of mRNA levels. BMC Genomics *14*, 185.

Podrabsky, J.E., and Culpepper, K.M. (2012). Cell cycle regulation during development and dormancy in embryos of the annual killifish Austrofundulus limnaeus. Cell Cycle *11*, 1697–1704.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. Bioinformatics *21* (*Suppl 1*), i351–i358.

Reichwald, K., Lauber, C., Nanda, I., Kirschner, J., Hartmann, N., Schories, S., Gausmann, U., Taudien, S., Schilhabel, M.B., Szafranski, K., et al. (2009). High tandem repeat content in the genome of the short-lived annual fish Nothobranchius furzeri: a new vertebrate model for aging research. Genome Biol. *10*, R16.

Rondeau, E.B., Messmer, A.M., Sanderson, D.S., Jantzen, S.G., von Schalburg, K.R., Minkley, D.R., Leong, J.S., Macdonald, G.M., Davidsen, A.E., Parker, W.A., et al. (2013). Genomics of sablefish (Anoplopoma fimbria): expressed genes, mitochondrial phylogeny, linkage map and identification of a putative sex gene. BMC Genomics *14*, 452.

Santoni, D., Castiglione, F., and Paci, P. (2013). Identifying correlations between chromosomal proximity of genes and distance of their products in protein-protein interaction networks of yeast. PLoS ONE *8*, e57707.

Shaw, W.M., Luo, S., Landis, J., Ashraf, J., and Murphy, C.T. (2007). The C. elegans TGF-beta Dauer pathway regulates longevity via insulin signaling. Curr. Biol. *17*, 1635–1645.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS ONE *6*, e21800.

Talavera, G., and Castresana, J. (2007). Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst. Biol. *56*, 564–577.

Terzibasi, E., Valenzano, D.R., and Cellerino, A. (2007). The short-lived fish Nothobranchius furzeri as a new model system for aging studies. Exp. Gerontol. *42*, 81–89.

Terzibasi, E., Valenzano, D.R., Benedetti, M., Roncaglia, P., Cattaneo, A., Domenici, L., and Cellerino, A. (2008). Large differences in aging phenotype between strains of the short-lived annual fish Nothobranchius furzeri. PLoS ONE *3*, e3866.

Thévenin, A., Ein-Dor, L., Ozery-Flato, M., and Shamir, R. (2014). Functional gene groups are concentrated within chromosomes, among chromosomes and in the nuclear space of the human genome. Nucleic Acids Res. *42*, 9854–9861.

Tozzini, E.T., Baumgart, M., Battistoni, G., and Cellerino, A. (2012). Adult neurogenesis in the short-lived teleost Nothobranchius furzeri: localization of neurogenic niches, molecular characterization and effects of aging. Aging Cell *11*, 241–251.

Tozzini, E.T., Dorn, A., Ng'oma, E., Polačik, M., Blažek, R., Reichwald, K., Petzold, A., Watters, B., Reichard, M., and Cellerino, A. (2013). Parallel evolution of senescence in annual fishes in response to extrinsic mortality. BMC Evol. Biol. *13*, 77.

Valdesalici, S., and Cellerino, A. (2003). Extremely short lifespan in the annual fish Nothobranchius furzeri. Proc. Biol. Sci. *270* (*Suppl 2*), S189–S191.

Valenzano, D.R., Kirschner, J., Kamber, R.A., Zhang, E., Weber, D., Cellerino, A., Englert, C., Platzer, M., Reichwald, K., and Brunet, A. (2009). Mapping loci associated with tail color and sex determination in the short-lived fish Nothobranchius furzeri. Genetics *183*, 1385–1395.

Valenzano, D.R., Sharp, S., and Brunet, A. (2011). Transposon-Mediated Transgenesis in the Short-Lived African Killifish Nothobranchius furzeri, a Vertebrate Model for Aging. G3 (Bethesda) *1*, 531–538.

Volff, J.N., Nanda, I., Schmid, M., and Schartl, M. (2007). Governing sex determination in fish: regulatory putsches and ephemeral dictators. Sex Dev. *1*, 85–99.

Wang, J., and Kim, S.K. (2003). Global analysis of dauer gene expression in Caenorhabditis elegans. Development *130*, 1621–1634.

Zahn, J.M., Sonu, R., Vogel, H., Crane, E., Mazan-Mamczarz, K., Rabkin, R., Davis, R.W., Becker, K.G., Owen, A.B., and Kim, S.K. (2006). Transcriptional profiling of aging in human muscle reveals a common aging signature. PLoS Genet. *2*, e115.

Zdobnov, E.M., and Apweiler, R. (2001). InterProScan–an integration platform for the signature-recognition methods in InterPro. Bioinformatics *17*, 847–848.

Zhang, J., Nielsen, R., and Yang, Z. (2005). Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol. Biol. Evol. *22*, 2472–2479.

Zullo, J.M., Demarco, I.A., Piqué-Regi, R., Gaffney, D.J., Epstein, C.B., Spooner, C.J., Luperchio, T.R., Bernstein, B.E., Pritchard, J.K., Reddy, K.L., and Singh, H. (2012). DNA sequence-dependent compartmentalization and silencing of chromatin at the nuclear lamina. Cell *149*, 1474–1487.

## Manuskript III: Outgroups and Positive Selection: The Nothobranchius furzeri Case

Trends in Genetics

CellPress

Forum

# Outgroups and Positive Selection: The *Nothobranchius furzeri* Case

Arne Sahm,[1] Matthias Platzer,[1] and Alessandro Cellerino[1,2,*]

**Applications of positive selection analysis increase with the number of species for which genome/transcriptome sequences become available. Using the recently sequenced turquoise killifish (*Nothobranchius furzeri*) genome as an example, we compare two different approaches based on different outgroup selection. The combination of these two methods allows the origin of positively selected sites in aging-related genes of the *N. furzeri* genome to be determined.**

The search for signatures of positive selection in protein-coding sequences has become a standard analysis to apply when new genomes or transcriptomes become available. In interspecies comparisons, protein-coding sequences under positive selection show more nonsynonymous substitutions than expected by random genetic drift, indicating that at least a proportion of those substitutions provided a selective advantage by altering gene function. Statistical models based on the ratio of nonsynonymous to synonymous substitution rate ($d_N/d_S$ ratio) are widely used in comparative genomics and have provided many insights into adaptive evolution (e.g., [1–3]). A link between specific genes – as well as sites within those genes – and the evolution of specific phenotypic traits can be postulated by comparing the coding sequences of a species (or a clade) showing the trait of interest with ortholog sequences of other, related species lacking that trait. The design of this taxonomic comparison is critical and, strictly speaking, an association between a given trait and the pattern of positive selection can be postulated only if a sister taxon (i.e., the most closely related species/clade) not showing the trait is available for analysis, which is a rather rare situation in genome-wide comparisons. Since genome papers are also read by researchers with cellular and molecular backgrounds not necessarily trained in evolutionary biology, it is useful to comment on the effects of outgroup selection on data analysis.

As a case in point, we illustrate here recently reported positive selection patterns in the genome of the annual fish *N. furzeri* [4,5]. This is the shortest-lived vertebrate that can be bred in captivity and is becoming an increasingly popular model organism since the effects of genetic and nongenetic interventions on aging and aging-associated phenotypes can be rapidly assessed [6,7] (see also Platzer and Englert in this issue). *N. furzeri* is adapted to an annual life cycle as it inhabits seasonal habitats (ponds) whose duration limits the lifespan of the adults, and the survival of embryos during the dry season is ensured by the ability to enter into a state of developmental arrest (diapause). Within the genus *Nothobranchius* there is considerable variation in captive lifespans and species originating from more humid habitats are longer lived [6]. *Aphyosemion* is the sister 'normal' (non-annual) genus to *Nothobranchius*. Here we define branch 1 the branch leading to the common ancestor of the two genera and branch 2 the branch leading to the common ancestor of all *Nothobranchius* spp. (Figure 1).

Recently two groups independently sequenced the genome of *N. furzeri* and applied two different strategies for the identification of genes under positive selection. The two analyses are prototypical of two different approaches to this problem. Their comparison provides insights of broad relevance for scholars interested in the genetic control of aging and life-history traits.

One study was performed by Valenzano *et al.* [5] and used as outgroups the available fish genomes and compared their protein-coding genes with the transcriptome of *N. furzeri*. This is a logical and broadly used strategy and has the advantage that the quality of the available genomes allows comparison of a larger number of genes and provides more power to detect positively selected genes. Through this analysis nearly 500 genes under positive selection were detected. The thrilling result of this analysis was that several genes clearly implicated in aging such as *BAX*, *FOXO1*, *IGF1R*, *INSR*, *IRS1*, *LMNA*, and *XRCC5* showed signs of positive selection, often in multiple sites. In addition, the *CEL* gene, under positive selection in the long-lived bowhead whale [8] and naked mole rat [9], was also under positive selection in *N. furzeri*. Due to the sparsity of sequenced fish genomes, the closest relative of *N. furzeri* that could be included in the analysis was the platyfish *Xiphophorus maculatus*. The two species, however, diverged between 70 and 50 million years ago [10] and therefore the origin for each of these positively selected sites along this long evolutionary branch could not be determined. It is likely that only a subset of these substitutions occurred in coincidence with the evolution of annualism or of short lifespan while others are more ancient evolutionary events.

The second study, by Reichwald *et al.* [4], took a complementary approach and assembled *de novo* the transcriptomes of five different *Nothobranchius* species, including *N. furzeri* sister species and *Aphyosemion striatum* from the nonannual sister genus. The latter inhabits permanent waters and its captive lifespan is of the order of years. This approach reduces the number of genes that can be analyzed (13 637 vs 11 748) but allows substitutions predating the evolution of

Trends in Genetics

CellPress

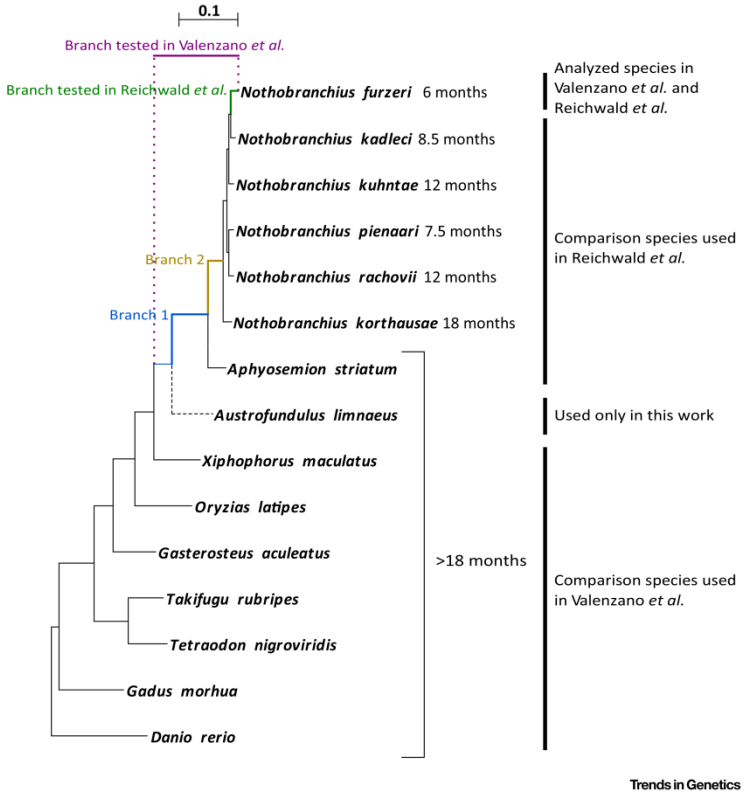



Figure 1. Nucleotide-Based Phylogenetic Tree of Species that Were Used in the Respective Genome-Scale Scans for Positively Selected Genes in the Works of Reichwald *et al.* [4] and Valenzano *et al.* [5]. Additionally, the position of the annual killifish *Austrofundulus limnaeus* is shown here. The time values next to the Latin species names report the median lifespan of each species [4]. The alignment is based on concatenation of 1399 genes. The represented tree is the consensus of 303 different trees created by splitting the alignment into fragments of 15 knt and calculating a tree for each fragment. The calibration bar refers to substitutions per nucleotide site.

annualism (branch 1), substitutions coinciding with the evolution of annualism (branch 2), and substitutions specific to *N. furzeri* as well as *Nothobranchius pienaari* (possibly related to very short lifespan) to be determined. Applying the latter approach we performed an analysis of the abovementioned genes (*BAX*, *CEL*, *FOXO1*, *IGF1R*, *INSR*, *IRS1*, *LMNA*, and *XRCC5*) with particular emphasis on IGF1R. The insulin/IGF pathway is the prototypic lifespan-regulating pathway in model organisms [11]; *IGF1R* was previously reported to be under positive selection in the long-lived Brandt's bat [12] and *IGF1R* variants are associated with longevity in humans [11]. Importantly, some of the detected substitutions were found in the ligand-binding or the kinase domain and could potentially alter the function of the receptor. *IGF1R* is duplicated in *N. furzeri* [3,4] and only one paralog (*IGF1RA*) is under positive selection [3]. We obtained the sequence of *IGF1RA* for *Nothobranchius korthausae* and an incomplete sequence from *A. striatum*; we also analyzed the public sequence of the South American killifish *Austrofundulus limnaeus* (GenBank: XP_013882332) and other more distant outgroups (Figure 1). Of the 15 reported substitutions [5], three were closely spaced in a region where our reference sequence matches both a third independently obtained sequence of *N. furzeri* (GenBank: HQ395756.1) and the outgroups (Figure 2). These substitutions are false positives due to an error in the gene model that was corrected in the current version of the genome (A. Brunet and D. Valenzano, personal communication). Of the remaining 12 substitutions, eight occurred in branch 1 or earlier, including a radical substitution in the ligand-binding domain (position 351: Val/Ile in the outgroups, Tyr in *N. furzeri*); the remaining four occurred in branch 2 and are therefore associated with the evolution of adaptations for annual life history (diapause) but also shorter lifespan, since all *Nothobranchius* spp. are shorter lived in nature than

Trends in Genetics

CellPress

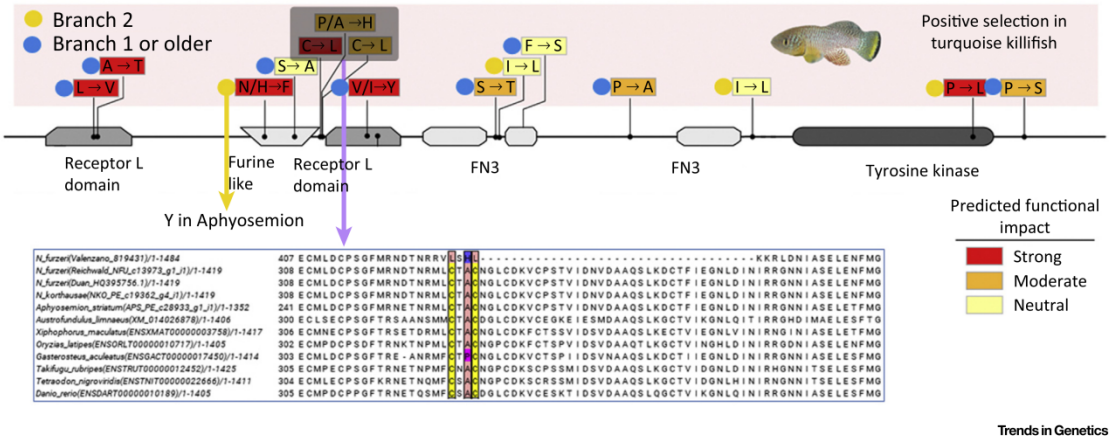



Figure 2. Pattern of Positive Selection on IGF1RA. The position of each positively selected site is indicated on a linearized protein scheme. Relevant protein domains are indicated at the bottom. The color within each box represents the strength of the predicted functional impact with respect to the residue observed in the outgroups of Valenzano *et al.* [5]. The box indicates an error in the gene model and the corresponding alignment is shown below. Please note that the IGFR1A sequences from Reichwald *et al.* [4] and Duan *et al.* (GenBank: HQ395756.1) align to outgroups in this region. Blue circles indicate substitutions corresponding to branch 1 or earlier (Figure 1) and are therefore not related to annual life cycle. Yellow circles indicate substitutions corresponding to branch 2 and therefore coinciding with the evolution of an annual life cycle. Redrawn from [5].

nonannual relatives. Of these four sites, three were conservative substitutions with respect to *A. striatum* (Y351F, I1008L, and I667L) with predicted neutral effects. The fourth change is located in the tyrosine kinase domain and showed a radical substitution (P1305L) with a predicted strong functional effect. This substitution is particularly interesting because it suggests that changes in IGF1RA function were relevant for the evolution of diapause, fast growth, and short lifespan associated with an annual life cycle. The substitutions in *CEL*, *INSRA*, *FOXO1*, and *LMNA* all occurred in branch 1. *BAX* and *IRS1* show three substitutions each and, for both genes, two occurred in branch 1 and one in branch 2. Particularly interesting is the pattern of positive selection on *XRCC5*, a gene involved in DNA repair [3]: of ten sites, six occurred in branch 1, two occurred in subgroups of *Nothobranchius* spp., and two (355F and 888G) are unique for *N. furzeri* and suggest that some fine-tuning of its function was specifically selected in this species. In conclusion, positive selection on *INSRA*, *CEL*, *FOXO1*, and *LMNA* predates the evolution of annualism, some of the positively selected residues in *BAX*, *IRS1*, *IGFR1A*, and *XRCC5* evolved in coincidence with annualism and are related to shorter lifespan and adaptations for annual life cycle such as diapause, and *XRCC5* shows residues possibly relevant for the extraordinary short lifespan of *N. furzeri*. This latter gene therefore may represent the most interesting candidate for functional studies.

In summary, we compared two approaches to detect positively selected genes in the *N. furzeri* genome maximizing sensitivity or specificity, respectively. The comparison of these two approaches allows the origin of positively selected sites in prominent aging-related genes of the *N. furzeri* genome to be determined. We have deposited the assembled and annotated transcriptomes of five *Nothobranchius* species and *A. striatum* to allow colleagues interested in the pattern of positive selection on other genes to make similar comparisons (European Nucleotide Archive, accession range: HADW01000000–HAEIJ01000000).

## Acknowledgments

The authors received funding from the Scuola Normale Superiore, the Leibniz Institute on Aging, and the German Research Foundation (DFG: PL 173/8-1).

[1]Leibniz Institute on Aging, Fritz Lipmann Institute, Jena, Germany
[2]Laboratory of Biology Bio@SNS, Scuola Normale Superiore, Pisa, Italy

*Correspondence: alessandro.cellerino@sns.it (A. Cellerino).

http://dx.doi.org/10.1016/j.tig.2016.06.002

## References

1. Davies, K.T. *et al.* (2015) Family wide molecular adaptations to underground life in African mole-rats revealed by phylogenomic analysis. *Mol. Biol. Evol.* 32, 3089–3107
2. Kosiol, C. *et al.* (2008) Patterns of positive selection in six mammalian genomes. *PLoS Genet.* 4, e1000144
3. Roux, J. *et al.* (2014) Patterns of positive selection in seven ant genomes. *Mol. Biol. Evol.* 31, 1661–1685
4. Reichwald, K. *et al.* (2015) Insights into sex chromosome evolution and aging from the genome of a short-lived fish. *Cell* 163, 1527–1538
5. Valenzano, D.R. *et al.* (2015) The African turquoise killifish genome provides insights into evolution and genetic architecture of lifespan. *Cell* 163, 1539–1554
6. Cellerino, A. *et al.* (2015) From the bush to the bench: the annual *Nothobranchius* fishes as a new model system in biology. *Biol. Rev. Camb. Philos. Soc.* 91, 511–533
7. Harel, I. *et al.* (2015) A platform for rapid exploration of aging and diseases in a naturally short-lived vertebrate. *Cell* 160, 1013–1026
8. Keane, M. *et al.* (2015) Insights into the evolution of longevity from the bowhead whale genome. *Cell Rep.* 10, 112–122
9. Kim, E.B. *et al.* (2011) Genome sequencing reveals insights into physiology and longevity of the naked mole rat. *Nature* 479, 223–227
10. Near, T.J. *et al.* (2012) Resolution of ray-finned fish phylogeny and timing of diversification. *Proc. Natl. Acad. Sci. U. S. A.* 109, 13698–13703
11. Fontana, L. *et al.* (2010) Extending healthy life span–from yeast to humans. *Science* 328, 321–326
12. Zhang, G. *et al.* (2013) Comparative analysis of bat genomes provides insight into the evolution of flight and immunity. *Science* 339, 456–460

## Manuskript IV: Parallel evolution of genes controlling mitonuclear balance in short-lived annual fishes.

*Aging Cell* (2017) **16**, pp488–496 Doi: 10.1111/acel.12577

# Parallel evolution of genes controlling mitonuclear balance in short-lived annual fishes

Arne Sahm,[1] Martin Bens,[1] Matthias Platzer[1] and Alessandro Cellerino[1,2]

[1]Leibniz Insitute on Ageing, Fritz-Lipmann Institute, Jena 07745, Germany
[2]Bio@SNS, Scuola Normale Superiore, Pisa 56124, Italy

## Summary

**The current molecular understanding of the aging process derives almost exclusively from the study of random or targeted single-gene mutations in highly inbred laboratory species, mostly invertebrates. Little information is available as to the genetic mechanisms responsible for natural lifespan variation and the evolution of lifespan, especially in vertebrates. Here, we investigated the pattern of positive selection in annual (i.e., short-lived) and nonannual (i.e., longer-lived) African killifishes to identify a genomic substrate for evolution of annual life history (and reduced lifespan). We identified genes under positive selection in all steps of mitochondrial biogenesis: mitochondrial (mt) DNA replication, transcription from mt promoters, processing and stabilization of mt RNAs, mt translation, assembly of respiratory chain complexes, and electron transport chain. Signs of paralleled evolution (i.e., evolution in more than one branch of *Nothobranchius* phylogeny) are observed in four out of five steps. Moreover, some genes under positive selection in *Nothobranchius* are under positive selection also in long-lived mammals such as bats and mole-rats. Complexes of the respiratory chain are formed in a coordinates multistep process where nuclearly and mitochondrially encoded components are assembled and inserted into the inner mitochondrial membrane. The coordination of this process is named mitonuclear balance, and experimental manipulations of mitonuclear balance can increase longevity of laboratory species. Our data strongly indicate that these genes are also casually linked to evolution lifespan in vertebrates.**

**Key words: evolution; gerontogenes; lifespan; longevity regulation; longevity gene; molecular biology of aging; mortality.**

## Introduction

The current molecular understanding of the aging process derives almost exclusively from the study of random or targeted single-gene mutations in highly inbred laboratory species, mostly invertebrates. Little information is available as to the genetic mechanisms responsible for natural lifespan variation and the evolution of longevity, especially in vertebrates. Yet, natural variability in lifespan across vertebrate species greatly exceeds the magnitude of life extension that has been obtained by single-gene manipulations, and a comparative approach may reveal novel genetic pathways that are responsible for evolution of lifespan.

Correspondence
*Alessandro Cellerino, Leibniz Insitute on Ageing, Fritz-Lipmann Institute, Jena 07745, Germany. Tel.: +39-050-3152756; fax: +39-050-3152760; e-mail: alessandro.cellerino@sns.it*

Accepted for publication *23 December 2016*

The increasing availability of sequenced genomes and transcriptomes of related species with differing lifespans can facilitate the identification of putative aging-related genes by analysis of positive selection. Positive selection is the evolutionary process by which a mutation becomes fixed in a population because it increases fitness. If two branches of an evolutionary tree differ in a key phenotype (lifespan, in this case), the genes under positive selection likely played a role in the evolution of that phenotype. In interspecies comparisons, positive selection on protein-coding sequences results in an increase in the rate of non-synonymous substitutions as compared with random genetic drift. Statistical models based on the ratio of non-synonymous to synonymous substitution rates ($d_N/d_S$) can identify specific amino acid codons within a given gene that evolved due to positive selection and are widely used in comparative genomics (Kosiol *et al.*, 2008; Roux *et al.*, 2014; Davies *et al.*, 2015).

One of the main limitations in applying this approach to the investigation of the genetic basis for lifespan evolution is the lack of a group of related species that are good laboratory organisms, are genetically tractable, and at the same time show naturally evolved large-scale differences in lifespan. Genome-wide scans for positive selection were performed in several long-lived mammals (bats, the naked mole-rat, the bowhead whale). However, it is not possible to establish a link between positively selected genes (PSGs) and evolution of longevity because the short-lived sister *taxon* (i.e., the most closely related species/clade) may not be available for analysis, making it impossible to exclude that of a codon change was selected before longevity evolved [for a discussion see (Sahm *et al.*, 2016a)] and it is very often impossible to relate a codon change to one of the several traits that distinguish two taxa (e.g., a PSG in *H. sapiens* may be related to longevity, bipedalism, absence of fur, speech, relative brain size, or any other trait that distinguish humans from apes).

Annual fishes of the genus *Nothobranchius* are small teleost fishes from East Africa adapted to the alternation of wet and rainy season. They inhabit ephemeral habitats that last a few months (Tozzini *et al.*, 2013). This short lifespan is retained under captive conditions and is coupled to rapid expression of a host of conserved age-associated phenotypes (Cellerino *et al.*, 2016). In addition, a key adaptation of annual fishes is the ability to enter diapause – a state when development halts – at specific stages during embryonic life, that is necessary to survive the dry season. The genus *Nothobranchius* evolved from a non-annual (therefore longer-lived) ancestor, the non-annual sister genus (Aphyosemion), is clearly identified (Furness *et al.*, 2015), and the two taxa provide a sharp phenotypic contrast. Duration of the habitat (aridity) strictly limits natural lifespan of Nothobranchius fishes.

We specifically tested whether differences in habitat duration led to the evolution of a different rate of senescence in Nothobranchius populations from southern and central Mozambique, a region characterized by a major gradient in aridity. Two independent evolutionary lineages of Nothobranchius are found in this area: *N. furzeri* and *N. kuhntae* belong to one lineage while *N. rachovii* and *N. pienaari* belong to another lineage (Dorn *et al.*, 2014). For each lineage, one species originates from semi-arid habitat (*N. furzeri* and *N. pienaari*, respectively) and another species from the humid habitat (*N. kuhntae* and *N. rachovii*, respectively). In both species pairs, the species from

more arid habitats showed shortened lifespan and accelerated expression of aging markers (Tozzini *et al.*, 2013), thereby providing a clear example of parallel evolution.

We previously sequenced and assembled the genome of *N. furzeri* as well as the transcriptomes of *N. kadleci* (the sister species of *N. furzeri*), *N. pienaari*, *N. rachovii*, and *N. kuhntae* together with *N. korthause* [a long-lived *Nothobranchius*, lifespan 18 months (Baumgart *et al.*, 2015)], and *Aphyosemion striatum* (lifespan > 3 years) as a representative of the non-annual sister genus (Reichwald *et al.*, 2015). We found seven genes under positive selection in *N. furzeri* and one in *N. pienaari*, another very-short-lived species, using the other six species of Nothobranchiidae as outgroups (Reichwald *et al.*, 2015). Here, we use a different selection of outgroups and analyze deeper branches of the *N. furzeri* phylogenetic tree to identify PSGs: (i) in coincidence with the evolution of annual life and (ii) showing parallel evolution in the two clades that are found in southern and central Mozambique.

Some results of this study were published in the form of preprint (Sahm *et al.*, 2016b).

## Results

### Genomewide scan strategy

In addition to sequence data presented previously (Reichwald *et al.*, 2015), we obtained from GenBank the RefSeq mRNA sequences of the phylogenetically closest outgroups from Ovalentaria (Fig. 1) and analyzed the pattern of positive selection along three internal branches of the tree: The first branch corresponds to the last common ancestor (LCA) of all *Nothobranchius* spp. (N-branch) and it marks the transition to annual life cycle. The other two branches correspond to the LCA of *N. pienaari* and *N. rachovii* (PR-branch) and LCA of *N. furzeri*, *N. kadleci* and *N. kuhntae* (FKK-branch), respectively. These two branches diverged in the Pleistocene, share the same distribution, and species belonging to the two clades can be found sympatric in the same pond (Dorn *et al.*, 2014). They represent therefore independent adaptations to the paleoclimatic changes of that period that was characterized by long-term progressive aridification of East Africa (Dorn *et al.*, 2014) and likely they were both subject to continued selection on adaptations linked to annual life cycle.

In each calculation, the background was the union of all the branches of the tree excluding the respective foreground branch, that is, when studying the N-branch the FKK and PR (and their child branches) are included in the background. In all comparisons, we defined PSGs based on nominal significant *P*-values (i.e., < 0.05, not corrected for multiple testing). This was a deliberate choice because of several reasons. First, we aim primarily at identifying parallel evolution at the level of pathway and not individual genes. Second, the number of genes strongly influences the sensitivity of Fisher's exact test, and it is not meaningful to perform GO analysis on lists containing few genes. Third, *de novo* transcriptome assembly projects inevitably generate incomplete data and a fraction of genes will show incomplete taxon coverage. We specifically tested whether PSGs have higher taxon coverage than the whole set of tested genes. However, this was not the case (Fig. 2) and only one PSG has a taxon coverage smaller than five.

### Positive selection acts on mitochondrial and mitonuclear balance proteins

We found 75 PSGs in the N-branch, 106 in the PR-branch, and 88 in the FKK-branch ($P < 0.05$, branch-site test; Tables S1–S3, Supporting information). Among these, four code for components of the mitochondrial respiratory chain complex I in the N-branch (GO:0005747, fold-enrichment = 14, $P = 0.0002$, Fisher's exact test; Fig. 3, Table S1, Supporting information). Therefore, emergence of annual life cycle is coincident with strong positive selection on mitochondrial respiration. This is in line with the evidence that diapause is linked to profound remodeling of mitochondrial physiology (Duerr & Podrabsky 2010). Three further genes of complex I are under positive selection in the PR-branch (fold-enrichment = 8.8, $P = 0.005$, Fisher's exact test) and one further gene in the FKK-branch, indicating parallel and continued positive selection on complex I during the evolutionary history of Nothobranchius (Fig. 3, Tables S2 and S3, Supporting information).

Strikingly, other nine genes were under positive selection in both the PR- and FKK-branches (Table 1). Among these, are *TFB2M* (transcription factor B2, mitochondrial) and *POLRMT* (polymerase (RNA) mitochondrial) that together with *TFAM* (transcription factor A, mitochondrial) form the ternary complex that transcribes the entire mitochondrial genome (Litonin *et al.* 2010) and *FASTKD5* (fast kinase domain 5) that is necessary for processing of mitochondrial mRNAs (Antonicka & Shoubridge 2015). Further signs of parallel positive selection were evident at the level of functional gene groups. In addition to *FASTKD5*, *FASTKD1*

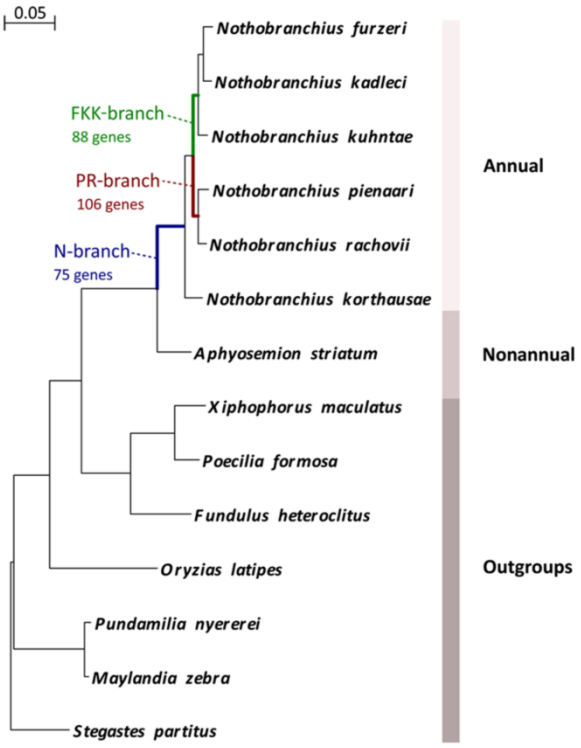



**Fig. 1** Phylogeny of the analyzed species. Maximum-likelihood tree figure. Phylogeny of the analyzed species and their life history. Maximum-likelihood nucleotide-based phylogenetic tree of species that were used for genome-scale scans for positively selected genes. Outgroups from Ovalentaria are indicated as well as the three branches (N-, PR-, and FKK-branch) that are reported in the text. The alignment is based on concatenation of 4865 genes. The represented tree is the consensus of 1046 different trees created by splitting the alignment in fragments of 15 knt and calculating a tree for each fragment. The calibration bar refers to substitutions per nucleotidic site.

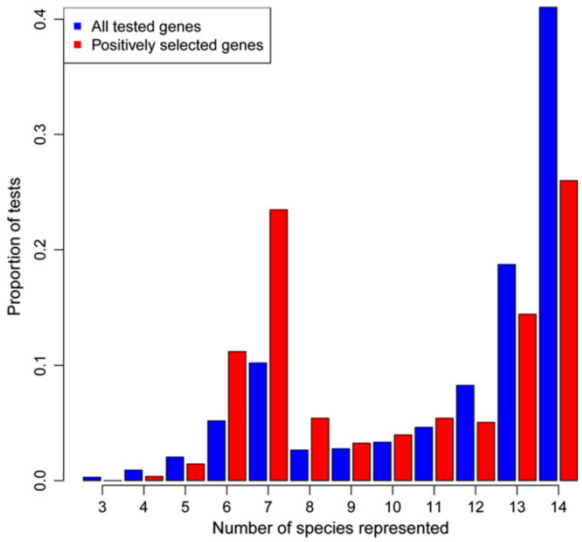



**Fig. 2** Distribution of taxon coverage for all tested genes (blue bars) and the positively selected genes (red bars). The *X*-axis reports number of taxa for which a gene sequence is available, and the *Y*-axis reports the fraction of genes falling into each coverage class.

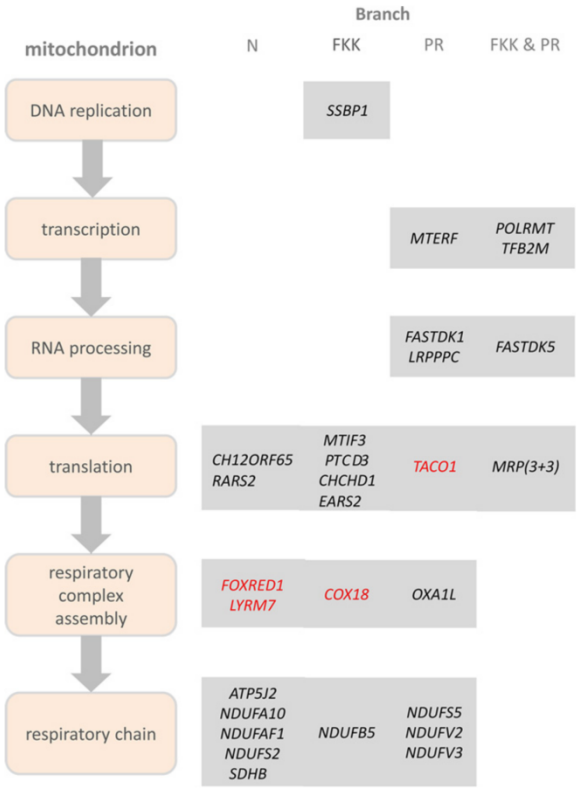



**Fig. 3** Genes controlling mitochondrial biogenesis and mitonuclear balance under selection in the three branches. Mitochondrial biogenesis was divided into the following processes: mtDNA replication, transcription from mitochondrial promoters, processing and stabilization of mitochondrial RNAs, translation, assembly of respiratory chain complexes and electron transport chain. Genes in black are classified based on their GO annotation genes in red genes are involved in mitochondrial biogenesis based on literature but not annotated as such in GO (see text for references). The term MRP indicates mitochondrial ribosomal proteins (MRPL53, MRPS31, and MPRS26 in FKK-branch and MRPL23, MRPL3, and MTG2 in the RP-branch, respectively).

and *LRPPPC* (leucine-rich pentatricopeptide repeat containing), that control stability of mitochondrial RNAs (Sasarman *et al.* 2010), were positively selected in PR-branch. Three mitochondrial ribosome proteins (MRPs) were under positive selection in each of the two branches (GO:0005761, fold-enrichment = 9.1 and 14.7, respectively, $P = 0.02$ and 0.01, Fisher's exact test for the PR- and FKK-branch, respectively). In addition, two recently identified MRPs (Koc *et al.* 2013) were positively selected in FKK-branch: *PTCD3* (pentatricopeptide repeat-containing protein 3) and *CHCHD1* (coiled-coil-helix-coiled-coil-helix domain containing protein 1). Two further genes important for translation of mitochondrial RNAs were also positively selected: *MTIF3* (mitochondrial translation initiation factor 3) in FKK-branch and *TACO1* (translational activator of mitochondrially encoded cytochrome C oxidase I) in PR-branch (Fig. 3).

Respiratory chain complexes are large protein complexes that undergo multistep assembly where nuclearly and mitochondrially encoded components are combined and inserted into the mitochondrial inner membrane (Ghezzi & Zeviani 2012). Several genes involved in this process were positively selected: *COX18* (cytochrome C oxidase assembly factor) (Sacconi *et al.* 2009) in FKK-branch, *OXA1*L (oxidase (cytochrome c) assembly 1-like) (Stiburek *et al.* 2007; Haque *et al.* 2010) in PR-branch, *FOXRED1* (FAD-dependent oxidoreductase domain containing 1; Fassone *et al.* 2010) and *LYRM7* (LYR motif containing 7) (Sanchez *et al.* 2013) in N-branch (Fig. 3). Therefore, proteins necessary for mitochondrial biogenesis and more specifically for expression of mitochondrially encoded genes and assembly of respiratory chain complexes show signs of parallel evolution. Altogether, among the observed 269 cases of positive selection along the three branches, 33 could be assigned to mitochondrial proteins and those involved in the mitochondrial biogenesis and mitonuclear balance (Fig. 3 and Tables S1–S3, Supporting information).

We also compared expression levels of genes in mitochondrial biogenesis and mitonuclear balance in two contrasts of a short- and a long-lived species: *N. furzeri* vs. *A. striatum* and mouse vs. naked mole-rat (Yu *et al.*, 2011). For the genes, 1-to-1 orthology relationships based on ENSEMBL IDs could be established for 23. Of these, 12 (*RARS2*, *FASTKD5*, *POLRMT*, *OXA1L*, *NDUFAF1*, *C12orf65*, *NDUFS2*, *MTG2*, *PTCD3*, *MRPS31*, *NDUFS5*, *NDUFB5*) have a lower expression in both long-lived species ($P$-value = 0.005985, binomial test, ¼, Table S4, Supporting information).

## Gene enrichment is not due to expression bias or incomplete lineage sorting

To ensure the statistical significance of our observation and exclude that biases due to the transcriptome assembly process and sequencing biases – in particular toward highly expressed genes – are responsible for the enrichment of mitochondrial proteins, we performed a simulation where we built two gene sets for each of three tested branches: an expression-adjusted background gene set and a "mitochondrial biogenesis" gene set. The later was derived from the union of the GO categories

**Table 1** Genes that are positively selected both in PR- and FKK-branch

| Gene symbol | Gene name | Function |
|---|---|---|
| *ETAA1* | Ewing tumor-associated antigen 1 | DNA damage response |
| *POLRMT* | Polymerase (RNA) mitochondrial (DNA directed) | Transcription of mtDNA |
| *PRRC2C* | Proline-rich coiled-coil 2C | Poly-A RNA binding |
| *APOA1* | Apolipoprotein A-I | High-density lipoprotein particle binding |
| *FASTKD5* | FAST kinase domains 5 | Regulation of mitochondrial RNA stability |
| *TAF1C* | TATA box-binding protein (TBP)-associated factor, RNA polymerase I, C, 110 kDa | Transcription of nuclear DNA |
| *TFB2M* | Transcription factor B2, mitochondrial | Transcription of mtDNA |
| *CLIP1A* (Nfu_g_1_008997) | CAP-GLY domain containing linker protein 1a | Unknown |
| *SI:DKEYP-77H1.4* (Nfu_g_1_001190) | Uncharacterized | |

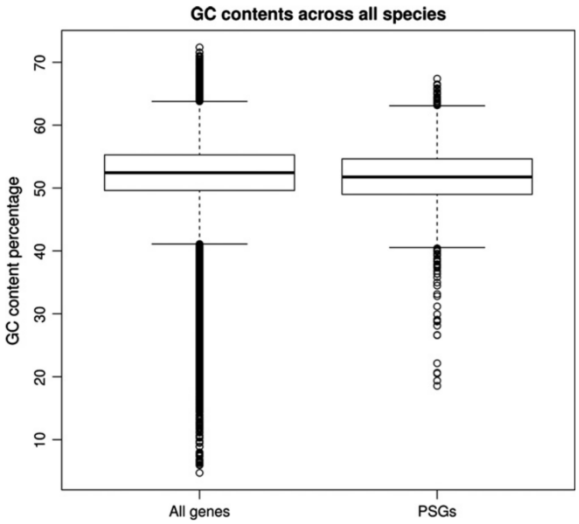



**Fig. 4** Distribution of GC content in all the tested transcripts and the positively selected genes. Data from all species and tests are pooled. Box plots indicate 10%, 25%, median, 75% and 90% quantiles, and points represent outliers.

mitochondrial RNA metabolic process (GO:0000959), mitochondrial translation (GO:0032543), cellular respiration (GO: 0045333), mitochondrial respiratory chain complex assembly (GO:0033108), mitochondrial morphogenesis (GO:0070584). Per simulation run, we then randomly draw for each of the three branches from the background set a number of genes that equals the number of PSGs that were identified in the respective branch and calculated the sum of drawn mitochondrial biogenesis genes. In none of 1.000.000 simulation runs, a higher number than 21 was observed (95% quantile: 11). We concluded that our finding of 33 cases of positive selection on mitochondrial biogenesis genes is highly significant (simulated $P < 10^{-6}$) and not caused by an expression or sequencing bias. Analysis of GC content demonstrated that PSGs did not differ from all analyzed genes (Fig. 4).

To exclude that enrichment was due to incomplete lineage sorting (Mendes & Hahn, 2016), we tested all PSGs that were identified initially based on a globally estimated tree again with a tree that was estimated using only the individually tested gene. Among all PSGs, 91% (244/269) were supported by this approach as well. Only one gene involved in mitochondrial biogenesis, namely *SDHB* in the N-branch, was not confirmed to be a PSG by this analysis.

### Positively selected genes are enriched among annualism-related genes

To derive independent evidence that the PSGs may be involved in the evolution of annual life style, we compared the union of the PSGs with two sets of differentially expressed genes (DEGs) in *N. furzeri*: (i) DEGs detected during brain aging (Baumgart *et al.*, 2014) and (ii) DEGs detected during diapause (Reichwald *et al.*, 2015). PSGs showed an over-representation among upregulated DEGs during diapause ($P = 0.023$, respectively, Fisher's exact test, Fig. 5) and among these 17 genes, *TFB2M* (PSG in both PR- and FKK-branch) and the assembly factor *NDUFA1* (PSG in the N-branch) are of relevance for mitonuclear balance. Over-representation of PSGs among upregulated DEGs is

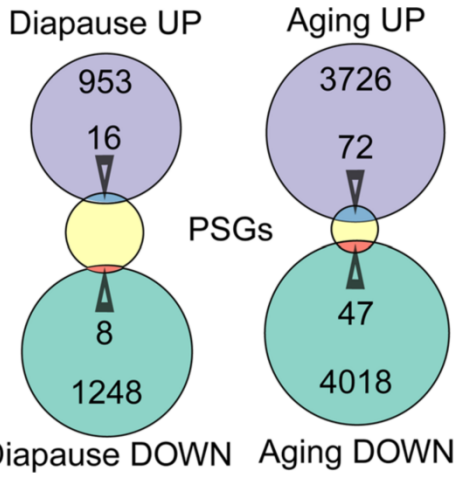



**Fig. 5** Overlap of positively selected genes (PSGs) with genes regulated during diapause or aging. Differentially expressed genes were obtained from Baumgart *et al.* (2014) and Reichwald *et al.* (2015) for brain aging and diapause, respectively. The arrowheads point to the intersection of the sets and indicate the number of genes in the respective intersections. The numbers within the circles indicate the number of genes in each set excluded from the intersections. The total number of PSGs in 267 in both cases.

observed also during aging ($P = 0.0093$, respectively, Fisher's exact test, Fig. 5), among these 47 genes, TFB2M is again present. PSGs upregulated during aging were also four genes of the cytokine–cytokine receptor interaction pathway (*CSF1RA*, *FLT1*, *IL2RGA*, *IL2ST*; dre04060 KEGG, $P = 0.0001$, Fisher's exact test).

In addition, we compared PSGs with results of longitudinal gene expression in *N. furzeri*. Gene co-expression network analysis revealed that *ETAA1*, positively selected in both PR- and FKK-branch, and *APOA1BP* (apolipoprotein A1 binding protein), the binding partner of

the PSG *APOA1,* are part of a module of co-regulated genes highly enriched with MRPs and complex I components and negatively correlated with longevity (Baumgart *et al.*, 2016; Fig. 6).

### Positively selected genes overlap with those in long-lived mammals

Previous analysis of PSGs in the *N. furzeri* genome suggested that some aging-relevant genes (e.g., the insulin like growth factor 1 receptor) can be positively selected both in short- and long-lived species (Valenzano *et al.*, 2015). We therefore compared *Nothobranchius* PSGs with PSGs detected by others in six species/clades of long-lived mammals (naked mole-rat, mole-rat LCA, blind mole-rat, human, bowhead whale and Brandt's bat). In all species, some PSGs overlap with those detected in *Nothobranchius* (Table S13, Supporting information). Of particular interest because under selection in more than two species/branches are: (i) *POLRMT*, that is a PSG in PR- and FKK-branch as well as in the Brandt's bat and the two extracellular matrix genes (ii) tenascin (*TNC*), a PSG in humans, mole-rats and in the FKK-branch regulated during both aging and diapause and (iii) Collagen type IV alpha 2 (*COL4A2*), a PSG in the naked mole-rat, the mole-rat LCA and in the FKK-branch also regulated during aging.

## Discussion

The coordinated synthesis and assembly of mitochondrially and nuclearly encoded components of the respiratory chain (mitonuclear balance) is a conserved longevity mechanism that is controlled by MRPs (Dillin *et al.*, 2002; Houtkooper *et al.*, 2013). Knock-down of MRPs during early life in *C. elegans* results in an impaired assembly of respiratory complexes and life extension. Studies in the mouse and *N. furzeri* have shown that MRPs and nuclearly encoded complex I components are tightly co-regulated and expression of these genes during early adult life is predictive of lifespan in vertebrates (Miwa *et al.*, 2014; Baumgart *et al.*, 2016). Further, inhibition of complex I activity during adult life prolongs lifespan and rejuvenates the transcriptome in *N. furzeri* (Baumgart *et al.*, 2016).

Here, we show that these same genes are under positive selection in annual fish strongly suggesting that that evolution of the genes controlling mitonuclear balance is causally linked to evolution of short lifespan and annual life cycle. This is supported by our findings that several of these genes are positively selected along more than one investigated branch and are differentially regulated during diapause and aging of the shortest lived *Nothobranchius* species.

At the single-gene level, of particular interest are PSGs that are detected both in the PR- and FKK-branch that represent examples of parallel evolution. *POLRMT* and *TFB2M* are part of the ternary complex that transcribes mitochondrial DNA (including mitochondrial rRNAs) and *TAF1C* (TATA box-binding protein-associated factor RNA polymerase I subunit C) that is part of the multisubunit SL1 complex, which is required for RNA polymerase I to synthesize ribosomal RNA. Therefore, these three genes are at the core of the process that controls the balance between biogenesis of cytosolic and mitochondrial ribosomes. *APOA1* (apolipoprotein A1) is a component of HDL particles that have an obvious relevance for human age-related diseases. Polymorphisms of *APOA1* are associated with coronary artery disease (Helgadottir *et al.*, 2016) and it is an interactor of the *APOE*, a well-described genetic risk factor for Alzheimer's and cardiovascular diseases (Mahley, 2016) and the locus with the largest statistical support for an association with extreme longevity (Broer *et al.*, 2015). Interestingly, its expression in the liver is correlated with body weight in mice (Pearson's correlation coefficient: −0.84 for females and −0.74 for males, http://phenome.jax.org/). *EAAT1* shows striking similarities with *APOA1*. Its expression in the liver correlates with female body weight in mice (Pearson's correlation coefficient: +0.94). *ETAA1* and *APOA1BP* have central positions in the gene module of co-expressed genes whose expression is negatively correlated with lifespan that also contains MRPs and complex I (Fig. 6), strongly suggesting that they are involved in mitonuclear balance. Interestingly, a function of *ETAA1* in DNA repair was recently demonstrated (Lee *et al.*, 2016) and the gene coding for another protein important in DNA repair, *XRCC5*, was previously shown to be under positive selection in *N. furzeri* (Valenzano *et al.*, 2015; Sahm *et al.*, 2016a). However, it is not possible to determine *in silico* whether the substitutions observed in the two lineages cause similar changes of mitochondrial function and parallel selection on the same genes does not represent a proof of functional convergence.

Are the genes under selection in short-lived species also involved in evolution of longevity? Data supporting this notion come from different

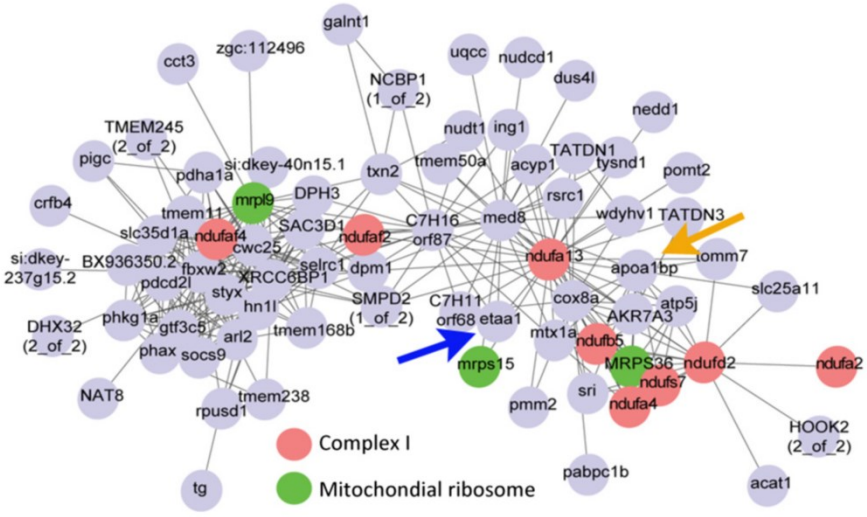



**Fig. 6** Position of ETAA1 and APOA1BP in the network of longevity-associated genes described by Baumgart *et al.* (2016). Picture reproduced with permission from Baumgart *et al.* (2016) and the gene annotation conforms to the annotation of transcripts described in Reichwald *et al.* (2015). Red genes code for complex I components and green genes for mitochondrial ribosome components.

studies of positive selection in the genomes of long-lived species. Ant workers can live on average ten times as long as their solitary ancestors and queens with 10 years at average and nearly 30 years at maximum even more than 100 times as long (Jemielity *et al.*, 2005). In an examination of seven ants genomes, highly significant enrichments of PSGs were documented for a series of GO terms that are related to the respiratory chain or mitochondrial biogenesis; especially mitochondrial electron transport (GO:0006120), mitochondrial respiratory chain complex I (GO:0005747), and mitochondrial large/small ribosomal subunit (GO:0005762/GO:0005763) (Roux *et al.*, 2014). The same study reported based on expression data obtained in the fire ant *S. invicata* that PSGs are highly expressed in queens, intermediately expressed in workers and weakest expressed in males which are the shortest lived ant caste. While the expression of PSGs correlates with lifespan of the respective caste, there is no differential expression across mitochondrial genes in general between queens and workers. This means that the association between PSGs and caste biased gene expression cannot be simply explained by higher overall levels of genes that are involved in mitochondrial activity but suggests a relation between queen-specific expression of PSGs and longer lifespan. Notably, consistent with the results of our study, there was no evidence found for positive selection on mitochondrial-encoded genes in ants. Furthermore, respiratory chain genes were found to be under positive selection in the bats *P. poliocephalus* and *M. lucifugus* (Shen *et al.*, 2010). Both are long-lived mammals, while *P. poliocephalus* reaches a maximum age of 23.6 years at a weight of 675 g resulting in a lifespan that is 1.7 times larger than expected based on the body mass, *M. lucifugus* even reaches a maximum age of 34 at a weight of only 10 g resulting in lifespan almost five times longer than expected based on body mass (Tacutu *et al.*, 2013).

Overlaps with long-lived mammals are detectable also at the level of single genes. Of particular interest is *POLRMT* this gene that codes for the mitochondrial RNA polymerase is a PSG in the PR- and FKK-branch and also in the Brandt's bat (Seim *et al.*, 2013) and, as discussed above, it is of key importance for mitonuclear balance. It is tempting to speculate that positive selection in short- and long-lived species modulates mitochondrial function in opposite directions. However, as discussed above, it is not possible to predict the functional impact of molecular evolution and this hypothesis will require experimental tests. Indirect evidence in favor comes from the observation that a significant fraction of mitochondrial biogenesis and mitonuclear balance genes are lower expressed in the longer lived element of two comparisons of long- and short-lived species: *N. furzeri* vs. *A. striatum* and mouse vs. naked mole-rat.

This hypothesis is also supported by direct measurements of complex I activity. Assays of mitochondrial physiology in the bivalve *Arctica islandica* (the longest lived metazoan with maximum lifespan exceeding 500 years) and two taxonomically related species of comparable size revealed lower activity of complex I resulting in reduced production of reactive oxygen species (Munro *et al.*, 2013). Similarly, low activity of complex I and low production of reactive oxygen species were related to longevity in homeotherm vertebrates (Brunet-Rossinni, 2004; Lambert *et al.*, 2010) and, finally, conditions that increase mouse longevity are associated with reduced expression of complex I (Miwa *et al.*, 2014).

Comparison of positive selection at the gene level between *Nothobranchius* and long-living mammals identified *TNC* and *COL4A2* as particularly interesting candidates as they are PSGs in two mammalian clades each and are also both differentially expressed in *Nothobranchius furzeri* aging (Reichwald *et al.*, 2015). These data lend further support to the notion that extracellular matrix genes are regulators of lifespan that derives from meta-analysis of genomewide transcript regulation (de Magalhaes *et al.* 2009), positive selection analysis (Li & de Magalhaes, 2013), and experimental approaches (Ewald *et al.*, 2015).

Finally, it should be noted that PSGs upregulated during aging were enriched for terms related to inflammation that are also known to be a highly conserved hallmark of aging at the transcriptome level (Baumgart *et al.*, 2014).

## Experimental procedures

### Genome-scale identification of positively selected genes

The basis for this work were protein-coding sequences (CDSs) of six *Nothobranchius* species (*N. furzeri*, *N. kadleci*, *N. kuhntae*, *N. pienaari*, *N. rachovii, and N. korthausae*) and *A. striatum* from transcriptome catalogs that were recently assembled and annotated (Reichwald *et al.*, 2015) with FRAMA (Bens *et al.*, 2016). The reads were adapter clipped with seqprep (https://github.com/jstjohn/SeqPrep) and quality trimmed with sickle (Joshi & Fass, 2011) before assembly [for more information about tissues, read numbers, filtered bases, etc., see Table S12 (Supporting information) or (Reichwald *et al.*, 2015)]. CDSs from seven additional outgroups (*Xiphophorus maculatus*, *Poecilia formosa*, *Fundulus heteroclitus*, *Maylandia zebra*, *Pundamilia nyererei*, *Stegastes partitus*, *Oryzias latipes*) were obtained from NCBI RefSeq (14.12.15) and assigned to ortholog groups by the best bidirectional BLAST hit criterion Camacho *et al.* (2009) against *N. furzeri*.

For each N. furzeri CDS isoform, the most similar isoform of each other species was determined by pairwise comparison. To reduce the risk of aligning nonhomologous codons, these sequences were required to have additionally at least a similarity of 70% with *N. furzeri* and 50% with each other species on protein level. The selected isoforms in each ortholog group were aligned with PRANK (Loytynoja & Goldman, 2008), which is the alignment software of choice for positive selection analysis (Fletcher & Yang, 2010). The alignments were stringently filtered with GBLOCKS (Talavera & Castresana, 2007) to remove gaps and unreliable alignment columns around them that could produce false signals of positive selection (−b2 = total number of sequences in the alignment, b4 = 30, t=c). Then, for each alignment the branch-site test of positive selection (Yang & Nielsen, 2002; Zhang *et al.*, 2005) was applied: The respectively tested branch (LCA, FKK, or PR) was marked as 'foreground', and all other branches were marked as 'background'. The program CODEML from the PAML (Yang, 2007) package was called separately for models M2a0 (model = 2, NSsites = 2; fix_omega = 1, omega = 1) and M2a (model = 2, NSsites = 2; fix_omega = 0, omega = 1) as described in the PAML User Guide (http://abacus.gene.ucl.ac.uk/software/pamlDOC.pdf). To calculate a *P*-value, the chi-square distribution with one degree of freedom was used to compare the likelihoods of both models: $P = \chi^2$ (2* (ln(likelihood(M2a))-ln(likelihood(M2a0))),1).

Sites under positive selection were inferred by the Bayes empirical Bayes method (Yang *et al.*, 2005) provided by CODEML. Sites that were predicted in a two amino acid frame next to a block which was deleted by GBLOCKS were removed and an adjusted *P*-value calculated. For 9017, 12028, and 10976 genes, *P*-values were calculated in the N-, FKK- and PR-branches, respectively (Table S5, Supporting information). We considered all genes with *P*-values $\leq 0.05$ as positively selected genes (PSGs).

Since high rates of false positive were detected in some automated genome-scale scans for PSGs in the past (Mallick *et al.*, 2009; Markova-Raina & Petrov, 2011), we demanded our candidates to fulfill further strict filter criteria. Candidates were removed that had: (I) not at least one species from the sister branch of the tested branch in the alignment,

for example, for the N-branch the presence of A. striatum was demanded, (II) less than four species in the alignment, (III) remained only few columns (<60 or <66.67%) or N. furzeri codons (<60%) of the alignment after GBLOCKS filtering, (IV) disproportional dN/dS ratios (e.g., ≥100 in foreground branch, >1 in background branch, <0.85 in foreground branch) were calculated by CODEML or (V) had an unreliably high fraction of inferred positively selected sites (more than 20%). Finally, we inspected all candidates on the FKK- and PR-branch manually as well as roughly ten percent of those on the LCA-branch and removed ten additional candidates (<5%) in total.

#### Phylogenetic tree

The phylogenetic tree that is needed for the analysis with CODEML was calculated based on the concatenated alignment of CDS isoforms of those 4865 genes with aligned isoforms from all species (Table S6, Supporting information). The final tree was the consensus of 1046 different trees created by splitting the alignment in fragments of 15 knt and calculating a tree for each fragment with DNAML from the PHYLIP (Felsenstein, 2005) package. All PSGs that were predicted with this globally estimated tree were again tested for positive selection with the same methods described above but with a tree that was estimated on the alignment of the respective gene. 65 of 75, 79 of 88, and 100 of 106 candidates were confirmed by this approach in the N-, FKK-, and N-branch, respectively (candidates that could not be confirmed are marked in Tables S1–S3, Supporting information).

#### Hypothesis-driven GO enrichments

We determined potential enrichments for the GO categories mitochondrial ribosome (GO:0005761) and mitochondrial respiratory chain complex I (GO:0005747) with Fisher's exact test. The set of tested genes that could be converted to entrez IDs served as background, that is, 7523, 9416, and 8745 genes for the N-, FKK-, and PR-branch, respectively (Table S7, Supporting information). As this was an hypothesis-driven approach, the *P*-values were not corrected for multiple testing.

#### Mitochondrial biogenesis enrichment simulation

For each of three tested branches, we built two gene sets, a background gene set and a mitochondrial biogenesis gene set. The background gene sets consisted of the tested genes of the respective branch that could be converted to Entrez IDs (https://www.ncbi.nlm.nih.gov/Entrez), that is, 7523, 9416, and 8745 genes for the N-, FKK-, and PR-branch, respectively (Table S7, Supporting information). To avoid biases due to expression differences between gene sets, we reduced the background sets to those genes within the 5–95% expression quantile of the union of PSGs across the three branches. This resulted in 6803, 8336, and 7882 genes, respectively (Tables S8 and S11, Supporting information). For the mitochondrial biogenesis gene sets, a union was built from the genes enlisted in the following five mitochondrial-related GO terms (GO:0000959, 0032543, 0045333, 0033108, 0070584). This union encompassed 331 genes (Table S9, Supporting information). For each branch, the mitochondrial biogenesis gene set consisted of the genes from this union that were also present in the background of the respective branch, resulting in 221, 250, and 245 genes, respectively (Table S9, Supporting information). In each simulation, round drawings were done for each branch from the respective background set and as often as PSGs were identified in that branch and could be converted to Entrez IDs, that is, 65, 73, 89 times (Table S10, Supporting information), respectively. Our simulation was conservative in the way that we did not reduce the number of drawings in each branch to the 5–95% expression quantile of the PSGs, giving the simulation a higher chance to draw genes from mitochondrial biogenesis set than we had in reality. At the end of each simulation round, it was counted how many drawn genes were in the mitochondrial biogenesis gene set for each branch and, finally, the sum across the three branches was calculated. One million simulation rounds were done.

## Author's contributions

AS and MB performed the analysis; MP and AC supervised the work; and AS, MP, and AC wrote the manuscript.

## Funding

This work was supported by the Leibniz Association (SAW-2012-FLI) and the German Research Foundation (DFG: PL 173/8-1) and a grant from Scuola Normale Superiore (CELLSNS2015).

## Conflict of interest

None declared.

## References

Antonicka H, Shoubridge EA (2015) Mitochondrial RNA granules are centers for posttranscriptional RNA processing and ribosome biogenesis. *Cell Rep.* doi:10.1016/j.celrep.2015.01.030.

Baumgart M, Groth M, Priebe S, Savino A, Testa G, Dix A, Ripa R, Spallotta F, Gaetano C, Ori M, Terzibasi Tozzini E, Guthke R, Platzer M, Cellerino A (2014) RNA-seq of the aging brain in the short-lived fish *N. furzeri* – conserved pathways and novel genes associated with neurogenesis. *Aging Cell* **13**, 965–974.

Baumgart M, Di Cicco E, Rossi G, Cellerino A, Tozzini ET (2015) Comparison of captive lifespan, age-associated liver neoplasias and age-dependent gene expression between two annual fish species: *Nothobranchius furzeri* and *Nothobranchius korthause*. *Biogerontology* **16**, 63–69.

Baumgart M, Priebe S, Groth M, Hartmann N, Menzel U, Pandolfini L, Koch P, Felder M, Ristow M, Englert C, Guthke R, Platzer M, Cellerino A (2016) Longitudinal RNA-Seq analysis of vertebrate aging identifies mitochondrial complex I as a small-molecule-sensitive modifier of lifespan. *Cell Syst.* **2**, 122–132.

Bens M, Sahm A, Groth M, Jahn N, Morhart M, Holtze S, Hildebrandt TB, Platzer M, Szafranski K (2016) FRAMA: from RNA-seq data to annotated mRNA assemblies. *BMC Genom.* **17**, 54.

Broer L, Buchman AS, Deelen J, Evans DS, Faul JD, Lunetta KL, Sebastiani P, Smith JA, Smith AV, Tanaka T, Yu L, Arnold AM, Aspelund T, Benjamin EJ, De Jager PL, Eirkisdottir G, Evans DA, Garcia ME, Hofman A, Kaplan RC, Kardia SL, Kiel DP, Oostra BA, Orwoll ES, Parimi N, Psaty BM, Rivadeneira F, Rotter JI, Seshadri S, Singleton A, Tiemeier H, Uitterlinden AG, Zhao W, Bandinelli S, Bennett DA, Ferrucci L, Gudnason V, Harris TB, Karasik D, Launer LJ, Perls TT, Slagboom PE, Tranah GJ, Weir DR, Newman AB, van Duijn CM, Murabito JM (2015) GWAS of longevity in CHARGE consortium confirms APOE and FOXO3 candidacy. *J. Gerontol. A Biol. Sci. Med. Sci.* **70**, 110–118.

Brunet-Rossinni AK (2004) Reduced free-radical production and extreme longevity in the little brown bat (*Myotis lucifugus*) versus two non-flying mammals. *Mech. Ageing Dev.* **125**, 11–20.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421.

Cellerino A, Valenzano DR, Reichard M (2016) From the bush to the bench: the annual *Nothobranchius* fishes as a new model system in biology. *Biol. Rev. Camb. Philos. Soc.* **91**, 511–533.

de Magalhaes JP, Curado J, Church GM (2009) Meta-analysis of age-related gene expression profiles identifies common signatures of aging. *Bioinformatics* **25**, 875–881.

Davies KT, Bennett NC, Tsagkogeorga G, Rossiter SJ, Faulkes CG (2015) Family wide molecular adaptations to underground life in African mole-rats revealed by phylogenomic analysis. *Mol. Biol. Evol.* **32**, 3089–3107.

Dillin A, Hsu AL, Arantes-Oliveira N, Lehrer-Graiwer J, Hsin H, Fraser AG, Kamath RS, Ahringer J, Kenyon C (2002) Rates of behavior and aging specified by mitochondrial function during development. *Science* **298**, 2398–2401.

Dorn A, Musilova Z, Platzer M, Reichwald K, Cellerino A (2014) The strange case of East African annual fishes: aridification correlates with diversification for a savannah aquatic group? *BMC Evol. Biol.* **14**, 210.

Duerr JM, Podrabsky JE (2010) Mitochondrial physiology of diapausing and developing embryos of the annual killifish Austrofundulus limnaeus: implications for extreme anoxia tolerance. *J. Comp. Physiol. B.* **180**, 991–1003.

Ewald CY, Landis JN, Porter Abate J, Murphy CT, Blackwell TK (2015) Dauer-independent insulin/IGF-1-signalling implicates collagen remodelling in longevity. *Nature* **519**, 97–101.

Fassone E, Duncan AJ, Taanman JW, Pagnamenta AT, Sadowski MI, Holand T, Qasim W, Rutland P, Calvo SE, Mootha VK, Bitner-Glindzicz M, Rahman S (2010) FOXRED1, encoding an FAD-dependent oxidoreductase complex-I-specific molecular chaperone, is mutated in infantile-onset mitochondrial encephalopathy. *Hum. Mol. Genet.* **19**, 4837–4847.

Felsenstein J (2005) PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Fletcher W, Yang Z (2010) The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. *Mol. Biol. Evol.* **27**, 2257–2267.

Furness AI, Reznick DN, Springer MS, Meredith RW (2015) Convergent evolution of alternative developmental trajectories associated with diapause in African and South American killifish. *Proc. Biol. Sci.* **282**, 20142189.

Ghezzi D, Zeviani M (2012) Assembly factors of human mitochondrial respiratory chain complexes: physiology and pathophysiology. *Adv. Exp. Med. Biol.* **748**, 65–106.

Haque ME, Elmore KB, Tripathy A, Koc H, Koc EC, Spremulli LL (2010) Properties of the C-terminal tail of human mitochondrial inner membrane protein Oxa1L and its interactions with mammalian mitochondrial ribosomes. *J. Biol. Chem.* **285**, 28353–28362.

Helgadottir A, Gretarsdottir S, Thorleifsson G, Hjartarson E, Sigurdsson A, Magnusdottir A, Jonasdottir A, Kristjansson H, Sulem P, Oddsson A, Sveinbjornsson G, Steinthorsdottir V, Rafnar T, Masson G, Jonsdottir I, Olafsson I, Eyjolfsson GI, Sigurdardottir O, Daneshpour MS, Khalili D, Azizi F, Swinkels DW, Kiemeney L, Quyyumi AA, Levey AI, Patel RS, Hayek SS, Gudmundsdottir IJ, Thorgeirsson G, Thorsteinsdottir U, Gudbjartsson DF, Holm H, Stefansson K (2016) Variants with large effects on blood lipids and the role of cholesterol and triglycerides in coronary disease. *Nat. Genet.* **48**, 634–639.

Houtkooper RH, Mouchiroud L, Ryu D, Moullan N, Katsyuba E, Knott G, Williams RW, Auwerx J (2013) Mitonuclear protein imbalance as a conserved longevity mechanism. *Nature* **497**, 451–457.

Jemielity S, Chapuisat M, Parker JD, Keller L (2005) Long live the queen: studying aging in social insects. *Age* **27**, 241–248.

Joshi NA, Fass JN (2011) Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files.

Koc EC, Cimen H, Kumcuoglu B, Abu N, Akpinar G, Haque ME, Spremulli LL, Koc H (2013) Identification and characterization of CHCHD1, AURKAIP1, and CRIF1 as new members of the mammalian mitochondrial ribosome. *Front. Physiol.* **4**, 183.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A (2008) Patterns of positive selection in six Mammalian genomes. *PLoS Genet.* **4**, e1000144.

Lambert AJ, Buckingham JA, Boysen HM, Brand MD (2010) Low complex I content explains the low hydrogen peroxide production rate of heart mitochondria from the long-lived pigeon, Columba livia. *Aging Cell* **9**, 78–91.

Lee YC, Zhou Q, Chen J, Yuan J (2016) RPA-binding protein ETAA1 is an ATR activator involved in DNA replication stress response. *Curr. Biol.* **26**, 3257–3268.

Li Y, de Magalhaes JP (2013) Accelerated protein evolution analysis reveals genes and pathways associated with the evolution of mammalian longevity. *Age* **35**, 301–314.

Litonin D, Sologub M, Shi Y, Savkina M, Anikin M, Falkenberg M, Gustafsson CM, Temiakov D (2010) Human mitochondrial transcription revisited: only TFAM and TFB2M are required for transcription of the mitochondrial genes in vitro. *J. Biol. Chem.* **285**, 18129–18133.

Loytynoja A, Goldman N (2008) Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. *Science* **320**, 1632–1635.

Mahley RW (2016) Apolipoprotein E: from cardiovascular disease to neurodegenerative disorders. *J. Mol. Med.* **94**, 739–746.

Mallick S, Gnerre S, Muller P, Reich D (2009) The difficulty of avoiding false positives in genome scans for natural selection. *Genome Res.* **19**, 922–933.

Markova-Raina P, Petrov D (2011) High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 *Drosophila* genomes. *Genome Res.* **21**, 863–874.

Mendes FK, Hahn MW (2016) Gene tree discordance causes apparent substitution rate variation. *Syst. Biol.* **65**, 711–721.

Miwa S, Jow H, Baty K, Johnson A, Czapiewski R, Saretzki G, Treumann A, von Zglinicki T (2014) Low abundance of the matrix arm of complex I in mitochondria predicts longevity in mice. *Nat. Commun.* **5**, 3837.

Munro D, Pichaud N, Paquin F, Kemeid V, Blier PU (2013) Low hydrogen peroxide production in mitochondria of the long-lived *Arctica islandica*: underlying mechanisms for slow aging. *Aging Cell* **12**, 584–592.

Reichwald K, Petzold A, Koch P, Downie BR, Hartmann N, Pietsch S, Baumgart M, Chalopin D, Felder M, Bens M, Sahm A, Szafranski K, Taudien S, Groth M, Arisi I, Weise A, Bhatt SS, Sharma V, Kraus JM, Schmid F, Priebe S, Liehr T, Gorlach M, Than ME, Hiller M, Kestler HA, Volff JN, Schartl M, Cellerino A, Englert C, Platzer M (2015) Insights into sex chromosome evolution and aging from the genome of a short-lived fish. *Cell* **163**, 1527–1538.

Roux J, Privman E, Moretti S, Daub JT, Robinson-Rechavi M, Keller L (2014) Patterns of positive selection in seven ant genomes. *Mol. Biol. Evol.* **31**, 1661–1685.

Sacconi S, Salviati L, Trevisson E (2009) Mutation analysis of COX18 in 29 patients with isolated cytochrome c oxidase deficiency. *J. Hum. Genet.* **54**, 419–421.

Sahm A, Platzer M, Cellerino A (2016a) Outgroups and positive selection: the *Nothobranchius furzeri* case. *Trends Genet.* **32**, 523–525.

Sahm A, Bens M, Platzer M, Cellerino A (2016b) Convergent evolution of genes controlling mitonuclear balance in annual fishes. *bioRxiv*. doi: https://doi.org/10.1101/055780

Sanchez E, Lobo T, Fox JL, Zeviani M, Winge DR, Fernandez-Vizarra E (2013) LYRM7/MZM1L is a UQCRFS1 chaperone involved in the last steps of mitochondrial Complex III assembly in human cells. *Biochim. Biophys. Acta* **1827**, 285–293.

Sasarman F, Brunel-Guitton C, Antonicka H, Wai T, Shoubridge EA, Consortium L (2010) LRPPRC and SLIRP interact in a ribonucleoprotein complex that regulates posttranscriptional gene expression in mitochondria. *Mol. Biol. Cell* **21**, 1315–1323.

Seim I, Fang X, Xiong Z, Lobanov AV, Huang Z, Ma S, Feng Y, Turanov AA, Zhu Y, Lenz TL, Gerashchenko, MV, Hee Fan, D, Yim, S, Yao, X, Jordan, D, Xiong, Y, Ma, Y, Lyapunov, AN, Chen, G, Kulakova, OI, Sun, Y, Lee, SG, Bronson, RT, Moskalev, AA, Sunyaev, SR, Zhang, G, Krogh, A, Wang, J and Gladyshev, VN (2013) Genome analysis reveals insights into physiology and longevity of the Brandt's bat *Myotis brandtii*. *Nat. Commun.* **4**, 2212.

Shen YY, Liang L, Zhu ZH, Zhou WP, Irwin DM, Zhang YP (2010) Adaptive evolution of energy metabolism genes and the origin of flight in bats. *Proc. Natl Acad. Sci. USA* **107**, 8666–8671.

Stiburek L, Fornuskova D, Wenchich L, Pejznochova M, Hansikova H, Zeman J (2007) Knockdown of human Oxa1l impairs the biogenesis of F1Fo-ATP synthase and NADH:ubiquinone oxidoreductase. *J. Mol. Biol.* **374**, 506–516.

Tacutu R, Craig T, Budovsky A, Wuttke D, Lehmann G, Taranukha D, Costa J, Fraifeld VE, de Magalhaes JP (2013) Human ageing genomic resources: integrated databases and tools for the biology and genetics of ageing. *Nucleic Acids Res.* **41**, D1027–D1033.

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *Syst. Biol.* **56**, 564–577.

Tozzini ET, Dorn A, Ng'oma E, Polacik M, Blazek R, Reichwald K, Petzold A, Watters B, Reichard M, Cellerino A (2013) Parallel evolution of senescence in annual fishes in response to extrinsic mortality. *BMC Evol. Biol.* **13**, 77.

Valenzano DR, Benayoun BA, Singh PP, Zhang E, Etter PD, Hu CK, Clement-Ziza M, Willemsen D, Cui R, Harel I, Machado BE, Yee MC, Sharp SC, Bustamante CD, Beyer A, Johnson EA, Brunet A (2015) The African turquoise killifish genome provides insights into evolution and genetic architecture of lifespan. *Cell* **163**, 1539–1554.

Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591.

Yang Z, Nielsen R (2002) Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* **19**, 908–917.

Yang Z, Wong WS, Nielsen R (2005) Bayes empirical bayes inference of amino acid sites under positive selection. *Mol. Biol. Evol.* **22**, 1107–1118.

Yu C, Li Y, Holmes A, Szafranski K, Faulkes CG, Coen CW, Buffenstein R, Platzer M, de Magalhaes JP, Church GM (2011) RNA sequencing reveals differential

expression of mitochondrial and oxidation reduction genes in the long-lived naked mole-rat when compared to mice. *PLoS ONE* **6**, e26729.
Zhang J, Nielsen R, Yang Z (2005) Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. *Mol. Biol. Evol.* **22**, 2472–2479.

## Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article.

**Table S1** Results of positive selection analysis in the N-branch, ranked by *P*-value.

**Table S2** Results of positive selection analysis in the PR-branch, ranked by *P*-value.

**Table S3** Results of positive selection analysis in the FKK-branch, ranked by *P*-value.

**Table S4** Comparison of fold-changes in expression of mitonuclear and mitochondrial biogenesis gene sets.

**Table S5** Overview of the tested genes.

**Table S6** List of genes used to construct the tree in Fig. 1.

**Table S7** List of background genes used for GO analysis.

**Table S8** List of genes used as background for the simulation experiment (5–95% expression quantile of the PSGs).

**Table S9** List of mitochondrial biogenesis genes within the background genes list (Table S7).

**Table S10** Entrez orthologs of PSGs.

**Table S11** Expression levels of all tested genes.

**Table S12** Statistics of read data for each library.

**Table S13** Complete list of *Nothobranchius* PSGs overlapping with PSGs in long-lived mammals.

## Manuskript V: Long-lived rodents reveal signatures of positive selection in genes associated with lifespan and eusociality

# Long-lived rodents reveal signatures of positive selection in genes associated with lifespan and eusociality

Arne Sahm[1*], Martin Bens[1], Karol Szafranski[1], Susanne Holtze[2], Marco Groth[1], Matthias Görlach[1], Cornelis Calkhoven[3], Christine Müller[3], Matthias Schwab[4], Hans A. Kestler[1,5], Alessandro Cellerino[1,6], Hynek Burda[7], Thomas Hildebrandt[2], Philip Dammann[7,8], Matthias Platzer[1]

[1] Leibniz Institute on Aging – Fritz Lipmann Institute, Jena, Germany.

[2] Department of Reproduction Management, Leibniz Institute for Zoo and Wildlife Research, Berlin, Germany.

[3] European Research Institute for the Biology of Ageing, University of Groningen, University Medical Centre Groningen Groningen, The Netherlands.

[4] Department of Neurology; Jena University Hospital-Friedrich Schiller University, Jena, Germany.

[5] Institute of Medical Systems Biology, Ulm University, Ulm, Germany.

[6] Laboratory of Biology Bio@SNS, Scuola Normale Superiore, Pisa, Italy.

[7] Department of General Zoology, Faculty of Biology, University of Duisburg-Essen, Essen, Germany.

[8] University Hospital, University of Duisburg-Essen, Essen, Germany.

[*] To whom correspondence should be addressed. Tel: +49 3641 656050; Fax: +49 3641 656255; Email: arne.sahm@leibniz-fli.de; Present Address: Arne Sahm, Genome Analysis, Leibniz Institute on Aging, Fritz Lipmann Institute, Jena, Thuringia, 07745, Germany.

## Abstract

The genetic mechanisms that determine lifespan are poorly understood. Most research has been done on short lived animals and it is unclear if these insights can be transferred to long-lived mammals like humans. Some African mole-rats (Bathyergidae) have life expectancies that are multiple times higher than similar sized and phylogenetically closely related rodents. This naturally occurring longevity dwarfs any effect achievable by intervention in vertebrates to extend lifespan. Therefore, we studied the pattern of positive selection in African mole-rats and other long-lived rodents by combining genomic and transcriptomic data. We obtained data from 17 species and systematically scanned six extant and five ancestral lineages leading to longevity for positively selected gene candidates (PSGs). The non-redundant set of 319 PSGs contains regulators of mTOR and is enriched in functional terms associated with (i) processes that are regulated by the mTOR pathway, e.g. translation, autophagy and mitochondrial biogenesis, (ii) the immune system and (iii) antioxidant defense. Analyzing gene expression of PSGs during aging, we found a significant pattern of down-regulation in the long-lived naked mole-rat and up-regulation in the short-lived rat, fitting the antagonistic pleiotropy theory of aging. We further analyzed the pattern of PSGs regarding the evolution of eusociality in African mole-rats and found it in line with a scenario that their last common ancestor already had a social predisposition. At the example of *TXN* and *TF*, we showed the potential functional relevance of the positively selected sites by homology modeling on the protein level, which may encourage experimental follow-up studies.

# Introduction

Most of the available information about the genetic mechanisms that govern lifespan and aging were obtained by studying single-gene mutations in invertebrates or short-lived, highly inbred vertebrate species. However, it is not clear whether insights about aging relevant genes and pathways gained from these species can be applied to long-lived species like human (Austad 2009). In addition, lifespan extensions under artificial laboratory conditions resulting from single gene mutations or other genetic, pharmacologic and/or lifestyle interventions are far smaller than natural variation of lifespan among species shaped by natural selection. Moreover, it is not clear to what extent genetic variation is responsible for intraspecific heritable differences in lifespan overlaps with the genetic architecture of lifespan macroevolution. As a case in point, maximum lifespan in captivity varies about two orders of magnitude and is positively correlated with body mass in vertebrates (Austad 2005; de Magalhaes, et al. 2007), but the two traits are negatively correlated within species, the most extreme example being dog breeds (Fan, et al. 2016). Therefore, comparative evolutionary approaches that search for genetic differences between closely related species that are short- and long-lived with respect to their body mass may reveal novel candidate genes and pathways or open new perspectives on known ones.

Rodents are an ideal taxon for such an approach. While the majority of species is short-lived, such as mice, rats and hamsters, there are long-lived exceptions, such as chinchillas, blind mole rats (BMR, *Spalax* sp.) and several African mole-rat species including the naked mole-rat (NMR, *Heterocephalus glaber*) (Tacutu, et al. 2013; Fushan, et al. 2015). Furthermore, genome and transcriptome sequences of short- and long-lived species are available and can be used for comparative analysis.

African mole-rats (family Bathyergidae) are subterranean rodents that feed from roots and tubers. The family comprises six genera; for five out of these, maximum lifespan records are available for at least one species. Notably, and in contrast to most other rodents, none of these species has a maximum lifespan of below ten years or below the predictions of the power-law that describes body mass/lifespan relationships in mammals (Fushan, et al. 2015). At the extreme of this distribution, Zambian mole-rats from the *Fukomys micklemi* clade (Van Daele, et al. 2007) (the best studied representative being the Ansell´s mole-rat *F. anselli*, AMR), the giant mole-rat (GMR, *Fukomys mechowii*) and NMR, have maximum lifespans of at least ca. 20, 22 and 31 years, respectively. These values are 212%, 194% and 368% with respect to the predicted lifespan based on their body mass ((Tacutu, et al. 2013), GMR percentage calculated with own lifespan data and same formula). In contrast, the established biomedical model organisms rat (*Rattus norvegicus*) and mouse (*Mus musculus*) have a maximum lifespan of 3.8 and 4 years, respectively, which is 32% and 51% of the predicted value. Remarkably, the greater cane rat (*Thryonomys swinderianus*) that is closely related to the African mole-rats reaches only 28% of the predicted maximum lifespan (Tacutu, et al. 2013) (Fig. 1).

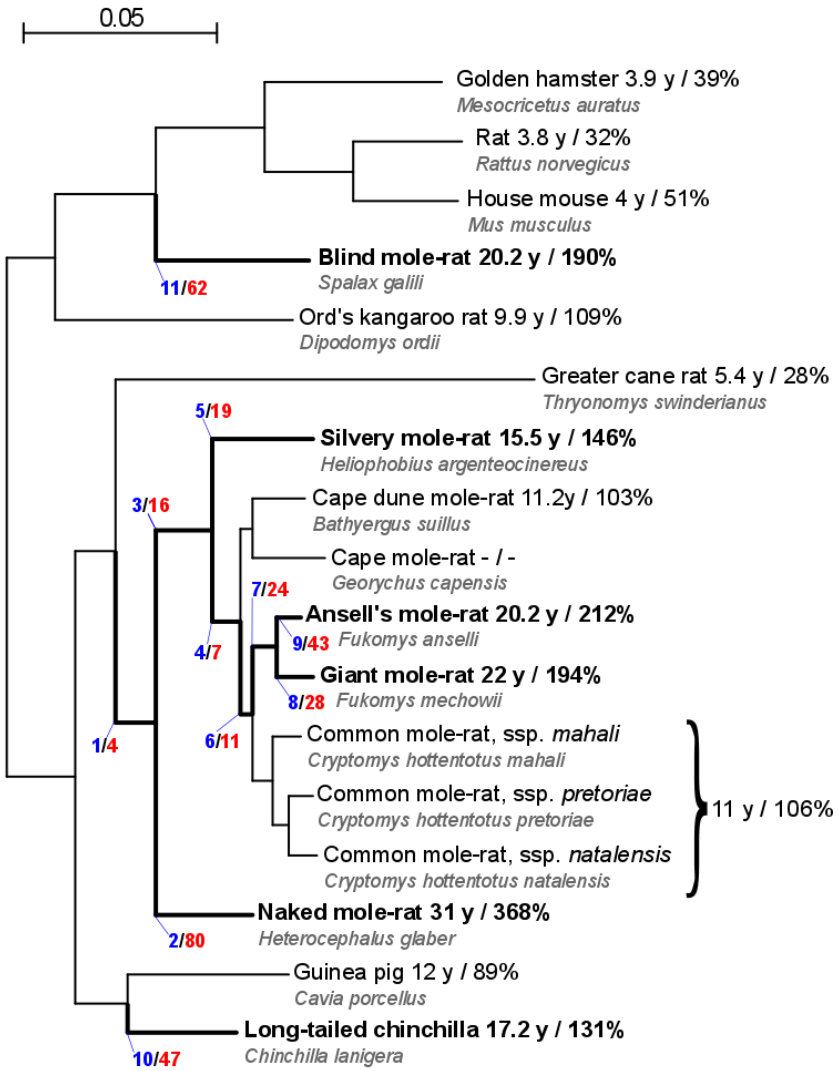



**Figure 1.** Nucleotide-based phylogeny of the analyzed rodents. Species or branches regarded in the present analyses as long-lived or leading to longevity, respectively, are depicted in bold. The branch numbers used in the text are shown in blue. The numbers of genes with signs of positive selection on the branches are colored in red. The first number after the species name shows the recorded maximum lifespan and the second number is the percentage of the observed vs. expected maximum lifespan based on the respective body mass. The maximum lifespans and ratios were taken from (Tacutu, et al. 2013), except for silvery mole-rat (personal communication by R. Sumbera) and giant mole rat (own data). For these two species, the expected maximum lifespans were calculated with the same mammalian allometric equation used by (Tacutu, et al. 2013). The scale bar represents 0.05 substitutions per site.

Due to a number of unique phenotypes, the NMR became the focus of intensive research (Gorbunova, et al. 2014). It was the first vertebrate for which eusociality was discovered ((Jarvis 1981)). The NMR shows (i) the longest lifespan among rodents, (ii) no aging-related decline in reproductive and physiological parameters, as well as (iii) no observable aging-related increase in mortality rate (Buffenstein 2008). Among thousands of examined animals only six recently discovered cases of spontaneous tumors have been described (Delaney, et al. 2016; Taylor, et al. 2017). Interestingly, cancer resistance is shared with BMR, which is also long-lived but, despite its name, rather distantly related to African mole-rats (Fig. 1). However, different mechanisms are proposed for cancer resistance in these two taxa. While high-mass hyaluronan mediated early contact inhibition was suggested as a key player in NMR (Seluanov, et al.

2009), a concerted necrotic cell death mechanism in response to hyperproliferation was proposed for BMR (Gorbunova, et al. 2012).

The search for signatures of positive selection represents a powerful approach to identify the genetic basis of these unique biological features. Positive selection is the fixation of an allele in a taxon driven by its positive effect on fitness. Once an adaptive phenotype evolved in a given species or evolutionary clade, some of the genes under positive selection likely play a role in it. In protein-coding sequences (CDSs), positive selection results in an increased rate of non-synonymous substitutions as compared to genetic drift. Statistical models based on the ratio of non-synonymous to synonymous substitution rates ($d_N/d_S$) are widely used in comparative genomics and allow the identification of specific amino acids within a given gene that changed due to positive selection (Kosiol, et al. 2008; Roux, et al. 2014; Sahm, Bens, Platzer and Cellerino 2017).

Consequently, several studies performed genome-scale scans for positively selected gene candidates (PSGs) in African mole-rats and BMR. The first study (Kim, et al. 2011) searched for PSGs on the very long NMR branch in a four-species comparison with human as an outgroup and the mouse and rat as further rodents. Among the 142 identified PSG candidates, three were members of a five-protein complex involved in alternative lengthening of the telomeres. The second study (Fang, Nevo, et al. 2014), used ten species with the guinea pig (*Cavia porcellus*) as most closely related species and scanned for PSGs along the branches leading to NMR, Damaraland mole-rat (*Fukomys damarensis*) and their last common ancestor (LCA), identifying 334, 179 and 82 candidates, respectively, including candidates associated with neurotransmission of pain in the NMR. A third study (Davies, et al. 2015) used species from all six African mole-rat genera and searched the branch of the LCA of all African mole-rats that follows divergence from the guinea pig. Signs of positive selection were identified in 513 genes, including loci associated with tumorigenesis, aging, morphological development and sociality. All three studies suffer from a methodological limitation that is common in positive selection studies: in none of these, a closer related species than guinea pig was included. As guinea pig is not the closest relative of African mole rats not expressing the phenotypes of interest, it cannot be excluded that fixation of detected signs of positive selection predates – and therefore could not contribute to – the evolution of these phenotypes (Sahm, et al. 2016). A fourth study (Fang, Seim, et al. 2014) examined the BMR branch using the Chinese hamster (*Cricetulus griseus*) as the most closely related outgroup. Among the 48 PSG candidates, several were linked to necrosis, inflammation and cancer.

To better resolve the above-mentioned ambiguities and to achieve a higher resolution of positive selection along rodent phylogenetic branches leading to longevity and eusociality, we analyzed genomic and transcriptomic data of 17 species – data from public sources and original data generated for this study. In particular, we generated genomic data for the greater cane rat as a key species absent from previous analysis and for the silvery mole-rat (SMR, *Heliophobius argenteocinereus*). We systematically scanned 11 evolutionary branches (6 corresponding to extant species and 5 to ancestral branches). This approach enables us to date precisely the occurrence of signatures of positive selection with respect to the evolution of the phenotypes of interest on multiple evolutionary branches of rodents. In addition, we generated RNA-seq data from young and old NMRs and laboratory rats (*Rattus norvegicus*) to analyze the overlap between PSGs and genes regulated during aging. Based on this, we discuss the implications of these results on our understanding of the genetic basis of aging, lifespan and sociality.

# Results and Discussion

As starting points for our analysis, we generated CDS libraries for five rodent species (NMR, AMR, GMR, SMR and greater cane rat) based on transcriptomic and genomic data. (Table S1/S2). Together with publicly available rodent CDS catalogs (Table S1), we obtained data for 17 species, including several additional African mole-rats, the chinchilla, BMR and short-lived outgroups like the guinea pig, mouse and rat (Fig. 1). From these sequences, we predicted orthologs and best matching isoforms between the species, calculated alignments and applied multiple times the branch-site test of positive selection (Zhang, et al. 2005).

Based on the lifespans of the extant species, we regarded six extant as well as five ancestral branches as leading to enhanced longevity and examined them for positive selection (Fig. 1). In total, we detected 341 PSGs ($p<0.05$, branch-site test). Our PSG assignment is based on nominal p-values, a common approach in genome-wide scans (Bakewell, et al. 2007; Gaya-Vidal and Alba 2014; Davies, et al. 2015) since the main error source of such analyses are alignment errors (Fletcher and Yang 2010) which result in extremely small p-values and therefore cannot be controlled by multiple test corrections. Furthermore, simulations have shown that the empirical false positive rate is very low if an appropriate filtering is used to remove alignment errors and unreliable results (Sahm, Bens, Platzer and Szafranski 2017).

Twenty genes were found on multiple branches (Table S3), resulting in a non-redundant set of 319 PSGs (Table S4-S15). Signs of positive selection for the same gene on multiple branches indicate possible parallel evolution. Among those, we found *AMHR2* (anti-Mullerian hormone receptor type 2) to be positively selected both on branch 2 (NMR) and branch 11 (BMR). While AMHR2 plays a role in male fetal development and in ovarian follicle development of the adult female (Durlinger, et al. 2002), no function with regard to aging is described yet. However, the protein kinase domain of AMHR2 contains the greatest number of longevity-selected positions based on a regression analysis with 33 mammalian species (Semeiks and Grishin 2012). This domain contains 3 of 8 and 2 of 3 positively selected sites on branch 2 (NMR) and branch 11 (BMR), respectively.

**Different studies on positive selection in mole-rats show minor overlaps**

First, we compared our list of PSGs with the PSGs detected in previous studies of positive selection in mole-rats (Kim, et al. 2011; Fang, Nevo, et al. 2014; Fang, Seim, et al. 2014; Davies, et al. 2015) (Table S16). As observed before, (Davies, et al. 2015) PSGs from different studies show no or small overlaps. This is not surprising because the branches examined in previous studies represent different phylogenetic entities than those used here, even though some of them are named similarly. For example, Kim et al. examined an “NMR branch” using the house mouse as closest related species (Kim, et al. 2011). In our study, the sister taxon to NMR is represented by other African mole-rats and the house mouse is used only as an outgroup (Fig. 1). In a similar way, the analysis of the African mole-rat ancestor by previous studies (Fang, Seim, et al. 2014; Davies, et al. 2015) differs from ours as we incorporated the greater cane rat as closest related short-lived species and used guinea pig as an outgroup. We therefore analyzed evolutionary processes on a shorter phylogenetic distance that closely matches the appearance of the phenotypes under investigation. In addition, there are methodological differences between the studies, e.g. regarding ortholog prediction or alignment filtering. Unfortunately,

the contribution of these technical variables to the discrepancies cannot be assessed as the alignments used for the previous studies are not available and cannot be compared with those generated and provided in our study (Supplement Data).

**Positive selection and age-related regulation are linked**

Next, we analyzed the direction of the regulation of PSGs during aging to identify potential links between positive selection on the analyzed branches and genetic determinants of lifespan. In general, directionality analysis of gene regulation during aging is complicated by the fact that the direction itself is not informative, whether the respective gene function is either causing or counteracting aging. E.g. up-regulation of a causative gene may accelerate aging and shorten lifespan while adaptive up-regulation to counteract aging phenotypes may extend longevity. We recently observed that PSGs in short-lived and fast-growing killifish were significantly more often up- than down-regulated during aging (Sahm, Bens, Platzer and Cellerino 2017). This finding is consistent with the concept of antagonistic pleiotropy (Hughes and Reynolds 2005) suggesting that the same genes that are positively selected in short-lived species for fast growth and maturation at young age are drivers of aging at old age. The antagonistic pleiotropy hypothesis is well supported, e.g. by the fact that growth rate and lifespan are negatively correlated, both between species and within many species (Bartke 2012; Fushan, et al. 2015). If, however, up-regulation of PSGs in short-lived species may cause aging, we hypothesized that selection for longevity is more compatible with attenuation of gene activity – either on the level of protein function or gene regulation – since avoiding damage is easier than improving repair.

Following this hypothesis, we performed RNA-seq in liver from old vs. young males of both long-lived NMRs and short-lived rats (>21 vs. 2-4 years and 24 vs. 6 months, respectively; Table S17-S19). Indeed, the union of PSGs showed preference for down-regulation in NMR and for up-regulation in rats in respect to all regulated genes ($p=0.0089$, Lancaster procedure (Dai, et al. 2014)). Moreover, the down-regulation in the long-lived and up-regulation in the short-lived species originate largely from the same genes as in a combined view on aging related expression changes in NMR and rat (Fig. 2), PSGs showed a highly significant preference for quadrant I (down in NMR, up in rat; $p=0.0014$, one-sided fisher test, quadrant I against the sum of II, III, IV). These results indicate that identified PSGs are associated with expression changes during aging of long- and short-lived rodents consistent with the antagonistic pleiotropy theory of aging.

To functionally annotate PSGs in respect to aging, we performed gene ontology (GO) term enrichment analysis. Regarding all genes, there was a significant enrichment for down-regulation in 126 terms during NMR aging while no term was enriched for up-regulation (Table S20, FDR<0.05, GAGE). The enriched 126 terms were summarized into 16 categories (Tables S21/S22, REVIGO). Among the six top categories are “translation” (GO:0006412), “cellular respiration” (GO:0045333), “response to oxidative stress” (GO:0006979) and “iron ion homeostasis” (GO:0055072) previously linked to aging (see below). With respect to possible pleiotropic effects, translation and cellular respiration are also key components of the growth program. To evaluate the PSGs in respect to these categories, we built the union of genes for each category and tested for overrepresentation of PSGs. Regarding all PSGs, there was a significant overlap with “cellular respiration” ($p=0.0022$, one-sided fisher test) and “response to oxidative stress” ($p=0.029$). Regarding only the 82 PSGs that were down-regulated in NMR and up-regulated during rat aging (quadrant I, Fig. 2), all four categories were significantly enriched (cellular respiration: $p=2.1*10^{-6}$,

response to oxidative stress: p=0.022, iron ion homeostasis: $p=8.5*10^{-4}$, translation: p=0.011; Table S23). This again suggests that PSGs are linked to aging relevant processes in an antagonistic pleiotropic way. The result is also consistent with the hyperfunction theory of aging that suggests that antagonistic pleiotropy works via a mechanism of "perverted" growth. According to this theory the growth program that is beneficial during youth is not entirely stopped after finishing development and causes damage from that point on. The theory further claims that the master regulator mTOR governs this growth program (Blagosklonny 2008, 2012).

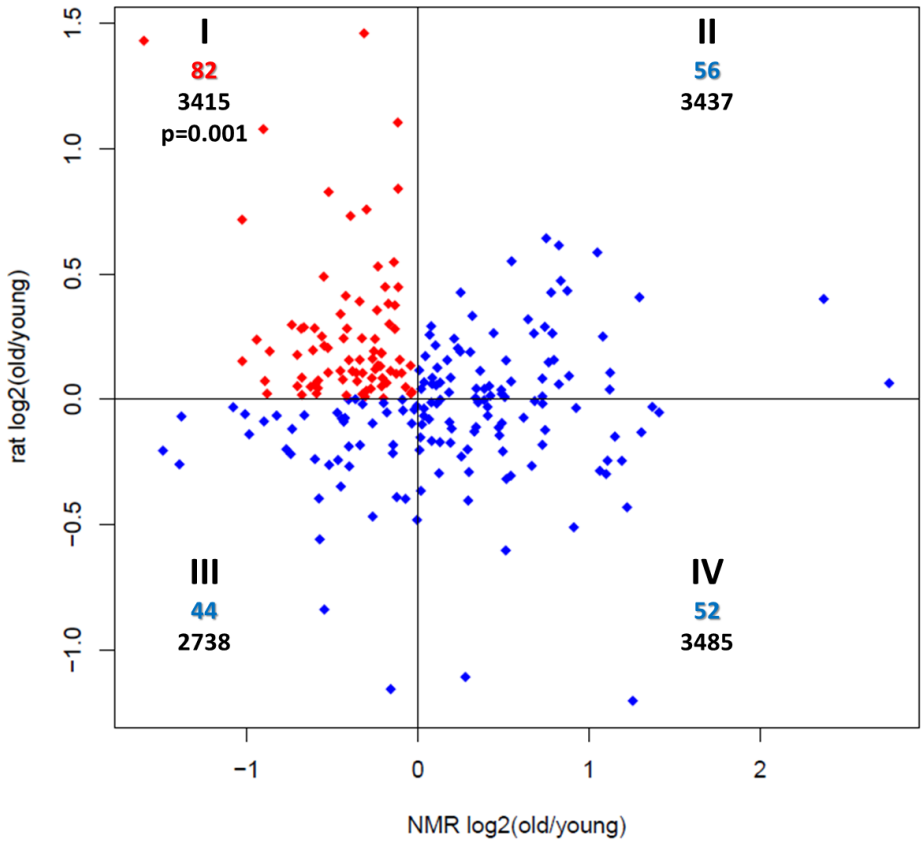



**Figure 2.** PSG expression changes during aging of NMR and laboratory rat. The roman numbers describe the quadrant, the colored numbers below that show the number of PSGs in the respective quadrant and the black numbers at the bottom give the total regulated genes in the quadrant. The red marked quadrant (I) represents PSGs that were down regulated in the long-lived NMR and up regulated in the short-lived rat and tested against the sum of three blue marked quadrants (II, III, IV) with Fisher's exact test (one-sided). The resulting p-value is shown in quadrant I. The total number of PSGs shown in this plot (234) is lower than the unique number of all PSGs (319) due to missing expression of genes in NMR and/or rat as well as missing log2-fold-changes in at least one of the species (DEseq2).

**Inflammation and host defense are enriched in branches leading to longevity**

Subsequently, we searched for enriched gene ontologies in the union of PSGs across the 11 branches along which longevity evolved and in each of these branches separately (Table S24). We found enrichments of genes involved in inflammatory response (GO:0006954; FDR=0.0068, Fisher's exact test) and defense response (GO:0006952, FDR=0.0092). Aging is tightly associated to the delicate balance between pro-inflammatory responses to resist potentially fatal infections and the inexorable damages

that are accumulated by this (Licastro, et al. 2005; Pitt and Kaeberlein 2015). Chronic inflammation is described as a major risk factor for aging and aging-related diseases such as atherosclerosis, diabetes, Alzheimer's disease, sarcopenia and cancer (Chung, et al. 2009).

**mTOR, autophagy and translation pathways show signs of positive selection leading to longevity**

On branch 2 (NMR), we found *RHEB* (Ras homolog enriched in brain) coding for a direct regulator of mTOR (mechanistic target of rapamycin) and on branch 9 (AMR) its paralog *RHEBL1* to be positively selected, a situation consistent with the concepts of parallel evolution as well as of subfunctionalization of genes after duplication. mTOR operates as a central regulator of cell metabolism, growth, inflammation and proliferation and was identified as a key regulator of aging and aging-related diseases in yeast, nematodes, fruit flies, and mice (Kenyon 2010; Johnson, et al. 2013).

mTOR is also a key regulator of autophagy (Jung, et al. 2010). Autophagy is a cellular protective cleaning mechanism, required for organelle homeostasis, especially mitochondria. While enhanced autophagy was shown to be associated with lifespan extension in worms, flies and mice, inhibition of autophagy, conversely, leads to premature aging in mice (Rubinsztein, et al. 2011). An essential autophagy gene, *LAMP2* (lysosomal associated membrane protein 2), was identified as PSG on branch 2 (NMR) and branch 11 (BMR). As a receptor for chaperone-mediated autophagy and a major protein component of the lysosomal membrane, LAMP2 is required for degradation of individual proteins through direct import into the lysosomal lumen (Cuervo and Dice 1996; Bandyopadhyay, et al. 2008). Aging-dependent decrease of *LAMP2* expression was observed in mouse liver. Reinstatement of juvenile LAMP2 levels in aged mice significantly reduces aging-dependent decline of cell function and restores the degree of cell damage to that found in young mice (Zhang and Cuervo 2008).

Besides the lysosome, another cellular protein quality control and degradation system is the proteasome. While impaired proteasome function and subsequent accumulation of misfolded proteins were tightly correlated with aging and aging-related neurodegenerative disorders like Parkinson’s and Alzheimer’s disease, long-lived humans have sustained proteasome activity (Chondrogianni, et al. 2000; Kevei and Hoppe 2014; Saez and Vilchez 2014). Two proteasome subunit genes, *PSMG1* (proteasome assembly chaperone 1) and *PSMB4* (proteasome subunit beta 4), were identified as PSGs on branch 11 (BMR). PSMB4 has been classified as a driver for several types of tumors (Lee, et al. 2014) and is a known interaction partner of PRP19 (pre-mRNA-processing factor 19 or senescence evasion factor) that is essential for cell survival and DNA repair (Beck, et al. 2008).

Another aging relevant downstream process regulated by mTOR is translation. We identified two ribosomal proteins, RPL7L1 and RPL27A, on branch 3 (LCA of all African mole-rats except NMR). While in general, cytosolic ribosomal proteins are up-regulated with aging in humans (Zahn, et al. 2006), rats (Ori, et al. 2015) and killifish (Reichwald, et al. 2015) both genes are significantly down-regulated during NMR aging (FDR≤0.05, DESeq2). This fits the down-regulation of translation-related processes during NMR aging in general (see above). Furthermore, the protein synthesis machinery is a driver of replicative senescence in yeast (Janssens, et al. 2015). The longitudinal aging study in killifish (Baumgart, et al. 2016) highlighted the starting values at 10 weeks and the amplitude of age-dependent increase of ribosomal proteins to be negatively correlated with lifespan. Inhibition of protein synthesis by reduction of

ribosomal proteins was shown to extend lifespan in worms (Hansen, et al. 2007) and mice (Hofmann, et al. 2015).

**Positive selection leading to longevity affects mitochondrial biogenesis and regulation of oxidative stress**

Besides regulation of cytoplasmic translation of nuclear encoded genes, mTOR is also involved in mitochondrial translation. There appears to be a complex interplay between mTOR signaling, mitochondrial gene expression and oxygen consumption as well as production of reactive oxygen species (ROS) (Schieke, et al. 2006; Bonawitz, et al. 2007; Bratic and Larsson 2013). Across multiple longevity-associated branches we identified PSGs that are involved in mitochondrial biogenesis (Table 1). We found, e.g., an enrichment of "mitochondrial translation" (GO:0032543, FDR=0.044) on branch 5 (SMR), the mitochondrial transcriptional termination factor (MTERF) on branch 2 (NMR) and six mitochondrial ribosomal proteins (MRPs) distributed on branches 5 (SMR), 7 (LCA of AMR and GMR) and 11 (BMR). Furthermore, nuclear encoded genes of respiratory chain complex I (*NDUFA9* and *NDUFB11*: NADH ubiquinone oxidoreductase subunits A9 and B11) and complex IV (*COX14*: cytochrome c oxidase assembly factor COX14) were identified as PSGs. Of note, we found 6 of these 15 genes to be significantly down-regulated during aging in NMR (Table 1). This suggests a functional relation of these genes to the aging process in an extremely long-lived rodent and is concordant with the down-regulation of genes involved in cellular respiration during NMR aging described above.

**Table 1.** Mitochondrial biogenesis genes under positive selection on longevity-associated branches.

| Gene | Positively selected on branch | | NMR aging | Function in mitochondrium |
|---|---|---|---|---|
| | Branch number | Branch description | | |
| *NDUFA9* | 8 | GMR | ↓ | Complex I |
| *MTERF* | 2 | NMR | ns | Transcription |
| *SDHAF2* | 2 | NMR | ns | Complex I |
| *UQCRC2* | 2<br>3 | NMR<br>LCA after NMR divergence | ↓ | Complex III |
| *NDUFB11* | 5 | SMR | ↓ | Complex I |
| *MRPS11* | 5 | SMR | ↓ | Translation |
| *MRPS36* | 5 | SMR | ↓ | Translation |
| *MRPL30* | 5 | SMR | ↓ | Translation |
| *TRNT1* | 11 | BMR | ns | RNA processing |
| *COX14* | 11 | BMR | ns | Complex IV |
| *DARS2* | 11 | BMR | ns | RNA processing |
| *MRPL57* | 11 | BMR | ns | Translation |
| *MRPL15* | 11 | BMR | ns | Translation |
| *MRPL28* | 7 | Fukomys LCA | ns | Translation |
| *GATC* | 10 | long-tailed chinchilla | ns | RNA processing |

Note: Branch numbers refer to figure 1. LCA – last common ancestor; ↓– significantly lower expressed in liver of old animals compared to young animals (FDR<0.05); ns – not significantly changed.

Studies in mouse and the short-lived killifish have shown that expression of MRPs and complex I genes is negatively correlated with individual lifespan (Miwa, et al. 2014; Baumgart, et al. 2016). Knock-down of MRPs in worms results in an impaired assembly of respiratory complexes and life-extension (Dillin, et al. 2002). Furthermore, we recently identified a significant enrichment of mitochondrial biogenesis genes including those for multiple MRPs, complex I components and MTERF among PSGs on two ancestral branches of annual killifishes on which lifespan was shortened considerably and independently from each other (Sahm, Bens, Platzer and Cellerino 2017). Altogether, these results raise again the intriguing possibility that similar or even the same genes could be causally linked to the evolution of both short and long lifespan.

Mitochondria are also the main source of ROS that cause oxidative stress, i.e. damages to DNA, proteins and other cellular components (Balaban, et al. 2005). Oxidative stress is thought to play a major role in the pathogenesis of neurodegenerative diseases (Kim, et al. 2015) and even the determination of lifespan in general ("oxidative stress theory of aging") (Barja 2014). On branch 3 (LCA of all African mole-rats except NMR), we found an enrichment of oxidoreductase activity (GO: GO:0016491; FDR=0.024) and positive selection of *TXN* (thioredoxin), coding for an oxidoreductase enzyme that acts as an antioxidant extending lifespan in fly (Umeda-Kameyama, et al. 2007) and potentially also in mice (Mitsui, et al. 2002; Perez, et al. 2011). As an example of continued evolution, *TXN* was found to be positively selected also on branch 7 (LCA of AMR and GMR). *SOD2* (superoxide dismutase 2) and *CCS* (copper chaperone for superoxide dismutase) are PSGs on branch 10 (chinchilla) and branch 2 (NMR), respectively. Both genes are involved in ROS defense and affect aging/lifespan in several species (Son, et al. 2009; Flynn and Melov 2013). This is interesting because in recent years, it has been repeatedly questioned that the

oxidative stress theory of aging has much relevance for bathyergid rodents, given that several studies failed to find improved antioxidant capacities and/or less accumulation of oxidative damage in NMRs compared to the much shorter-lived mice (Andziak, et al. 2005; Andziak and Buffenstein 2006; Andziak, et al. 2006). This is consistent with our finding of down-regulation of processes involved in "response to oxidative stress" (GO:0006979) during aging in NMR (see above). On the other hand, significantly higher levels of oxidative damage on proteins and lipids in non-reproductive as compared to reproductive females of the Damaraland mole-rat were found (Schmidt, et al. 2014). Since non-reproductive individuals live shorter (and hence age faster) than their reproductive counterparts in *Fukomys* sp. (Dammann and Burda 2006; Dammann, et al. 2011; Schmidt, et al. 2013), these results are consistent with the oxidative stress theory of aging. The diverse signs of positive selection on branch 2 (NMR), 3 (LCA of all African mole-rats except NMR) and 7 (LCA of AMR and GMR) may suggest that the impact of oxidative stress on aging differs between NMR and other African mole-rats.

ROS production and ROS-induced damage to biomolecules are intertwined with the formation of advanced glycation end-products (AGEs). AGEs are stable bonds between carbohydrates and proteins/lipids which are formed in a non-enzymatic fashion. AGEs activate membrane-bound or soluble AGER (AGE specific receptor) and AGEs/AGER have been linked to several aging-related diseases including Alzheimer's disease and diabetes (Vistoli, et al. 2013). Interestingly, *AGER* was found to be a PSG on branch 9 (AMR) and branch 10 (chinchilla). The role of AGEs/AGER in aging is complex and Janus-faced (Simm, et al. 2015). AGER is significantly up-regulated in liver during NMR aging. Similarly, in skin AGE levels rise with chronological age in AMR, but surprisingly are higher in the skin of slow aging breeders than of faster aging non-breeders (Dammann, et al. 2012)

**Additional links between positive selection and longevity**

The gene *APOA1* (apolipoprotein A1) was found as PSG on branch 7 (LCA of AMR and GMR) and significantly up-regulated during NMR aging in liver. APOA1 is a component of HDL-particles which are as transporter of cholesterol relevant for aging-associated diseases. Polymorphisms of APOA1 are associated to coronary artery disease (Helgadottir, et al. 2016). Furthermore, APOA1 is an interaction partner of APOE a well-described genetic risk factor for Alzheimer's and cardiovascular diseases (Mahley 2016) and the locus with the largest statistical support for an association with extreme longevity (Broer, et al. 2015). Same as *MTERF* (see above), *APOA1* is one of nine genes that we recently found to be positively selected on both of two ancestral sister branches of annual fishes on which lifespan was independently reduced (Sahm, Bens, Platzer and Cellerino 2017).

*TF* (transferrin) was identified as PSG on branch 4 (LCA of Cape, Cape dune, giant, AMR and common mole-rats). TF is an iron-binding protein responsible for transport of iron in the bloodstream and therefore essential for iron homeostasis (Macedo and de Sousa 2008). Neurons regulate iron intake via the TF receptor and dysregulation of this tightly controlled process in the brain is associated with neurodegenerative, age-related diseases like Parkinson's and Alzheimer's (Hare, et al. 2013). TF is significantly down-regulated during NMR aging which is consistent with the down-regulation of "iron ion homeostasis" (GO:0055072) related processes during NMR aging in general (see above).

**Selection signatures of social evolution among African mole rats is consistent with a scenario of ancestral eusociality**

Although all African mole-rats are strictly subterranean and occupy similar nutritional niches, intra-familiar variety of social and mating systems is amazingly high. Solitariness and polygamy in some genera (*Heliophobius*, *Georychus* and *Bathyergus*) contrast sharply with social organization in others (*Heterocephalus*, *Fukomys* and *Cryptomys*). In the latter, stable monogamous bonding of (typically one) reproductive founder pair coupled with prolonged philopatry and reproductive altruism of their offspring result in extended and cooperatively breeding family units, which can grow to considerable size. There has been much debate whether eusociality in African mole-rats is a derived or ancestral trait (Jarvis and Bennett 1993; Burda, et al. 2000). The first scenario assumes a solitary LCA of African mole-rats (branch 1) and subsequent independent, parallel evolution of eusocial habits on branch 2 (NMR) and branch 6 (LCA of AMR, GMR and common mole-rats) and/or branch 7 (LCA of AMR and GMR) (Davies, et al. 2015). In contrast, the second scenario suggests a eusocial LCA of all mole-rats (branch 1) and independent loss of this phenotype in the SMR (branch 5) and the LCA of the genera *Bathyergus* and *Georychus* (Cape and Cape dune mole-rat). Recent phylogenetic approaches are more supportive of scenario 2 (Smorkatcheva and Lukhtanov 2014; Sobrero, et al. 2014) however, a PSG-based analysis of the issue is still lacking. Accordingly, we searched our data in support of one or the other scenario.

All four PSGs found on the ancestral branch 1 of all African mole-rats are involved in signaling, resulting in enrichments, e.g., of "positive regulation of cell communication" (GO:0010647, FDR=0.0092) and "positive regulation of signaling" (GO:0023056, FDR=0.0092). Enrichments in signal transduction were identified as major common pattern of the multiple independent occurrences of eusociality in bees (Woodard, et al. 2011) strongly suggesting parallel evolution in eusocial insects and mammals.

On the other hand, members of the major histocompatibility complex (MHC) were identified as PSG on branch 2 (NMR) and on branch 7 (LCA of AMR and GMR). MHC genes have a central function in the acquired immune system and immune dysfunction is involved in many neurodevelopmental disorders as well as social behavior deficits in mice and humans (Estes and McAllister 2015; Malkki 2016). MHCs have been implicated in language impairment and schizophrenia (Kodavali, et al. 2014; Nudel, et al. 2014). A human allele of *HLA-A* was associated with autism defined as a pattern of behavior identified by deficits in communication and reciprocal social interactions (Torres, et al. 2006).

Additionally, we found two signals with a potential link to social evolution on branch 2 (NMR), but not branch 6 (LCA of AMR, GMR and common mole-rats) or 7 (LCA of AMR and GMR). The first were innate immune system related enrichments "positive regulation of T cell activation" (GO:0050870, FDR=0.027) and "positive regulation of leukocyte cell-cell adhesion" (GO:1903039, FDR=0.027). In social ants and bees, it was shown that several innate immune genes have a pattern of accelerated amino acid evolution compared both to non-immunity genes in the same species and immune genes in solitary fly (Viljakainen, et al. 2009). Second, we found *NOTCH2* positively selected only on branch 2 (NMR). The encoded protein is one of four Notch-receptors. Notch signaling regulates interactions between physically adjacent cells and has a central role in the development of many tissues, including neurons (Guruharsha, et al. 2012). It was demonstrated that Notch signaling represses reproduction in worker honeybees depending on the presence of the queen and that chemical inhibition of Notch signaling can overcome the repressive effect of queen pheromone in regard to the worker ovary activity (Duncan, et al. 2016).

On the other hand, *HSD11B1* (hydroxysteroid 11-beta dehydrogenase 1) was identified as PSG on branch 6 (LCA of AMR, GMR and common mole-rats), but not branch 2 (NMR). The encoded protein catalyzes reversibly the conversion of the stress hormone cortisol to the inactive metabolite cortisone (Frick, et al. 2004). Cortisol concentration was shown to inversely correlate with social ranks in NMRs (Clarke and Faulkes 1997) and with anti-social and isolation behavior in human adolescents (Sanchez-Martin, et al. 2001; Hawes, et al. 2009). Furthermore, cortisol regulates carbohydrate metabolism that is another common enriched GO-Term in the evolution of eusociality in bees (Woodard, et al. 2011). Not linked to eusociality but still noteworthy, *HSD11B1* is significantly down-regulated during aging in the NMR and knockout of *HSD11B1* in mice improves their cognitive performance in aging (Holmes, et al. 2001). Furthermore, inhibition was described as a risk factor for cardiovascular disease and diabetes type 2 (Anderson and Walker 2013).

Taken together, our data are on best agreement with a scenario assuming eusociality (or a predisposition for it) in the LCA of all African mole-rats, followed by further independent, branch-specific evolution or loss of the phenotype leading to the distinct social genera that live today.

**Homology modeling suggests functional consequences of amino acid changes under positive selection**

To evaluate the structural impact of positively selected amino acid changes, we performed homology modeling using exemplary the sites of highest probability for selection in cytoplasmic thioredoxin (TXN) and transferrin (TF). As mentioned above *TXN* is positively selected on branch 7 (LCA of AMR and GMR). In TXN of these species, there is a tyrosine residue that replaces Cys69. The latter, together with Cys62 and Cys73, constitute highly conserved mammalian non-catalytic cysteines. The local structure around Cys69 and Cys62 in TXN is important for interaction with the cytoplasmic thioredoxin reductase (TR1; (Fritz-Wolf, et al. 2011)), which 'recycles', i.e. re-reduces, the catalytic cysteines of oxidized TXN. The modeling using the structure of a fully reduced human TXN (1ERT*;* (Weichsel, et al. 1996)) as template suggests that the rather bulky side chain of Tyr69 can be accommodated in the structure of TXN (Fig. 3A), hence allowing for a productive helical interface region to TR1. TXN recycling is inhibited by formation of a disulfide bridge between Cys62 and Cys69 (Fig S1; see (Hwang, et al. 2015)), e.g. under highly oxidative conditions, thereby diminishing the pool of catalytically active TXN under oxidative stress (Watson, et al. 2003; Hashemy and Holmgren 2008). Obviously, that disulfide bridge cannot form in AMR and GMR because of Cys69Tyr. From this, we conclude that Tyr69 is compatible with TXN recycling also under oxidative stress. Moreover, Cys69 is known to be a target for posttranslational modifications with impact on e.g. anti-apoptotic/apoptotic signaling pathways (for a review see: (Wu, et al. 2011)), raising interesting questions on physiological consequences of the Cys69Tyr replacement.

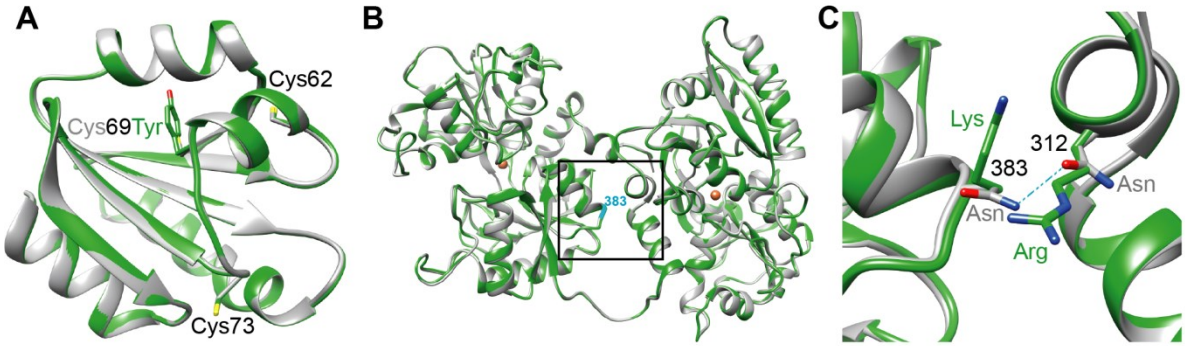



**Figure 3**. Homology models of Ansell's mole-rat (AMR) thioredoxin (TXN) and transferrin (TF). (A) Overview of the modeled AMR TXN structure (green) superimposed onto the fully reduced human TXN template structure (1ERT, grey). Residues discussed in the text are labeled, numbering according to position in the human sequence. Color code of Cys69 and Tyr69 corresponds to the respective ribbon representation. Heteroatoms: sulfur in yellow, oxygen in red. (B) Overview of the modeled AMR TF structure (green) superimposed onto the rabbit TF template structure (1JNF, grey). The position of the Asn383Lys site discussed in the text at the boxed center of the lobe interface numbered and indicated in cyan. Brown spheres: $Fe^{3+}$ coordinated in the template structure (1JNF, ion radius enlarged for better visibility). (C) Detail of the TF lobe interface. Shown is a magnification of the boxed region in (B). Coloring and numbering as in (B), side chain nitrogen atoms (blue), oxygen atoms (red). Potential hydrogen bond in 1JNF (light blue) as discussed in the text. Numbering (black) according to positions in the rabbit TF structure (1JNF).

TF is a PSG on branch 4 (LCA of Cape, Cape dune, giant, AMR and common mole-rats) and Ser383Lys is the site of highest probability for selection. Serum TFs form a bilobal structure, and each lobe contains two dissimilar domains with a single iron-binding site. Inspecting the structure of the AMR TF modeled on the rabbit protein (1JNF; (Hall, et al. 2002)) as template, we realized that Lys 383 is located at the interface between the two lobes (Fig. 3B). In the rabbit TF two juxtapositioned Asn residues at position 383 and 312 might form an H-bond and this constellation could stabilize the inter-lobe interactions (Fig. 3C). In contrast, the juxtaposition of the positively charged side chains of Lys383 and a conserved Arg312 in the AMR structural model (Fig. 3C) would be expected to weaken the lobe-lobe interaction due to electrostatic repulsion. The functional consequences for TF or TXN implied by this modeling require experimental validation.

## Conclusions

We provided a systematic scan for PSGs on evolutionary branches of the African mole-rat family and other rodents leading to longevity and eusociality. Due to the incorporation of species from all six genera of the African mole-rats as well as its closest relative, the cane rat, into the analysis, we were able to examine considerably more extant and ancestral branches than previous studies. This enabled the analysis to provide a high resolution of positive selection on branches on which the mentioned traits had evolved.

Analyzing the gene expression of PSGs, we found a highly significant pattern of down-regulation in the long-lived NMR and up-regulation in the short-lived rat, fitting the antagonistic pleiotropy theory of

aging (Medawar 1952) and the hyperfunction theory of aging. The latter claims mTOR as a central hub affecting aging and lifespan (Blagosklonny 2012). Correspondingly, the PSGs and enriched functional terms cover many of the processes that are regulated by the mTOR pathway, e.g. translation, autophagy and mitochondrial biogenesis. Furthermore, with *RHEB* and its ortholog *RHEB1L* direct regulators of mTOR (Groenewoud and Zwartkruis 2013) are under positive selection in two of the branches. In addition, we linked positive selection with immune system and the antioxidant defense, processes known to be involved in regulation of lifespan.

With regard to evolution of eusociality, our findings are in line with the theory that the LCA of all African mole-rats had at least a predisposition for social lifestyle that was lost in some lineages, while in other lineages the ancestral phenotype has further evolved, leading to the distinct social phenotypes in the extant species.

Moreover, we exemplarily showed potential functional relevance of the positively selected sites by homology modeling on the protein level. This may encourage experimental follow-up studies since all sequences and alignments including the identified positively selected sites are accessible via supplement data.

# Methods

**CDS data**

We examined nine African mole-rat species covering all six genera. Additionally, our analysis comprises eight outgroup species, including the long-lived BMR and the chinchilla. mRNA sequences of seven distantly related outgroup species were obtained from RefSeq along with their CDS annotation (Table S1). For the NMR we used a recently published *de novo* transcriptome assembly (Bens, et al. 2016). RNA-seq data for six mole-rat species was obtained from GenBank Sequence Read Archive, study SRP061925 (Davies, et al. 2015). The reads were assembled and annotated using FRAMA as described in (Bens, et al. 2016).

For AMR and GMR, purification of RNA from 13 and 17 tissues, respectively, was done using Qiagen RNeasy Mini Kit following the manufacturer's description. Novel RNA-seq was performed for both species as described in Table S2. *De novo* transcriptome assemblies of the generated data were performed using FRAMA (Bens, et al. 2016) (see Table S1).

For the SMR and the greater cane rat genome sequencing was performed to complement the transcriptome data. DNA was isolated from liver tissue of two female SMR individuals and a male liver of the greater cane rat using DNeasy Blood & Tissue (Qiagen). DNA was then converted to Illumina libraries and sequencing was done as given in Table S2. Sequence reads were cleaned by removal of adaptors and low-quality regions at the ends (i.e. regions with more than 10% with quality score ≤ 20). Low quality reads (i.e. less than 50% remained) and duplicons were discarded. *De novo* genome sequence assembly was performed using CLC assembler (CLC Genomics) with default settings. The CDS annotation was done using AUGUSTUS (Stanke, et al. 2006) with AMR CDSs as hint.

All animals were housed and euthanized compliant with national and state regulations. Read data was deposited as ENA (European Nucleotide Archive) study PRJEB20584.

**Identification of positively selected genes**

To scan on a genome-wide scale for genes under positive selection, we fed the CDSs of the described species set along with the branches we wanted to examine (Fig. 1) into the PosiGene pipeline (Sahm, Bens, Platzer and Szafranski 2017). GMR was used as PosiGene's anchor species. Regarding the SMR, for which we had both a genome and a transcriptome assembly, we used generally the transcriptome assembly, except for those ortholog groups in which no SMR ortholog could be assigned via transcriptomic but via genomic data. This was accomplished by calling the three PosiGene modules separately, feeding both assemblies independently in the first module (ortholog assignment) and deleting all genome-based SMR sequence in those ortholog groups that contained transcriptome-based SMR CDSs before calling the second module. An overview about the number of genes and sequences tested for positive selection in the different branches is shown in Table S1. We considered all genes with nominal p-values ≤ 0.05 as PSGs.

**Gene ontologies**

We determined enrichments for GO categories with Fisher's exact test based on on the R package GOstats. The resulting p-values were corrected using the Benjamini-Hochberg method (FDR).

**Differentially expressed genes during NMR and rat aging**

The young and old rats (strain Wistar) had an age of 6 (n=4) and 24 (n=5) months, respectively. The young NMRs had an age of 3.42±0.58 years (average±sd, n=6). The old NMRs were at least 21 years old (recorded lifetime in captivity, n=3). All examined animals were males. All animals were housed and euthanized compliant with national and state regulations. For both species, purification of RNA from liver samples was done using Qiagen RNeasy Mini Kit following the manufacturer's description. In short, we performed RNA-seq using Illumina HiSeq 2500 with 50 nt single read technology and a sequencing depth of at least 20 mio reads/sample (Table S17). For NMR, the read mapping was performed with STAR (Dobin, et al. 2013) (--outFilterMismatchNoverLmax 0.06 --outFilterMatchNminOverLread 0.9 --outFilterMultimapNmax 1) against the public genome (Bioproject: PRJNA72441) that we had annotated before by aligning the above mentioned NMR transcriptome reference using BLAT (Kent 2002) and SPLIGN (Kapustin, et al. 2008). Rat reads were aligned against the mentioned RefSeq reference using bwa aln (Li and Durbin 2009) (-n 2 -o 0 -e 0 -O 1000 -E 1000). Read data and counts were deposited as GEO (Gene Expression Omnibus) series GSE98746. Differentially expressed genes (FDR≤0.05, Table S18, S19) and fold-changes were determined with DESeq2 (Love, et al. 2014). GAGE (Luo, et al. 2009) was used to determine enriched gene ontologies based on fold-changes (Table S20). Gene ontologies with FDR≤0.05 were summarized using REVIGO (allowed similarity=0.5) (Supek, et al. 2011). Four of the six largest summarized categories of the resulting treemap (Table S21/S22) were further analyzed due their aging relevance (representative terms given): "translation" (GO:0006412), "cellular respiration" (GO:0045333), "response to oxidative stress" (GO:0006979) and "iron ion homeostasis" (GO:0055072). For each of these categories the union of genes across gene ontology terms was built. These unions were tested for significant overlaps with (i) the union of PSGs across branches and (ii) the union of PSGs across

branches that were down-regulated during aging in NMR and up-regulated in rat (Fisher's exact test). Functional annotation of the PSGs in respect to the four categories is given in Table S23).

**Homology modeling of protein structure**

Models were built in SWISS-MODEL (http://swissmodel.expasy.org;) (Arnold, et al. 2006; Biasini, et al. 2014). No further optimization was applied to the resulting TXN and TF models. Superimposition of the model and template structures and rendering was carried out using CHIMERA (Pettersen, et al. 2004).

## Acknowledgement

We thank Ivonne Görlich, Christiane Vole and Yoshiyuki Henning for excellent assistance, Debra Weih for proofreading the manuscript and Christoph Kaether for helpful discussions. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, PL 173/8-1 and DA 992/3-1), the European Community's Seventh Framework Programme (FP7-HEALTH-2012-279281) as well as the Leibniz association (SAW-2012-FLI-2).

## References

Anderson A, Walker BR. 2013. 11beta-HSD1 inhibitors for the treatment of type 2 diabetes and cardiovascular disease. Drugs 73:1385-1393.
Andziak B, Buffenstein R. 2006. Disparate patterns of age-related changes in lipid peroxidation in long-lived naked mole-rats and shorter-lived mice. Aging Cell 5:525-532.
Andziak B, O'Connor TP, Buffenstein R. 2005. Antioxidants do not explain the disparate longevity between mice and the longest-living rodent, the naked mole-rat. Mech Ageing Dev 126:1206-1212.
Andziak B, O'Connor TP, Qi W, DeWaal EM, Pierce A, Chaudhuri AR, Van Remmen H, Buffenstein R. 2006. High oxidative damage levels in the longest-living rodent, the naked mole-rat. Aging Cell 5:463-471.
Arnold K, Bordoli L, Kopp J, Schwede T. 2006. The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinformatics 22:195-201.
Austad SN. 2009. Comparative biology of aging. J Gerontol A Biol Sci Med Sci 64:199-201.
Austad SN. 2005. Diverse aging rates in metazoans: targets for functional genomics. Mech Ageing Dev 126:43-49.
Bakewell MA, Shi P, Zhang J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc Natl Acad Sci U S A 104:7489-7494.
Balaban RS, Nemoto S, Finkel T. 2005. Mitochondria, oxidants, and aging. Cell 120:483-495.
Bandyopadhyay U, Kaushik S, Varticovski L, Cuervo AM. 2008. The chaperone-mediated autophagy receptor organizes in dynamic protein complexes at the lysosomal membrane. Mol Cell Biol 28:5747-5763.
Barja G. 2014. The mitochondrial free radical theory of aging. Prog Mol Biol Transl Sci 127:1-27.
Bartke A. 2012. Healthy aging: is smaller better? - a mini-review. Gerontology 58:337-343.
Baumgart M, Priebe S, Groth M, Hartmann N, Menzel U, Pandolfini L, Koch P, Felder M, Ristow M, Englert C, et al. 2016. Longitudinal RNA-Seq Analysis of Vertebrate Aging Identifies Mitochondrial Complex I as a Small-Molecule-Sensitive Modifier of Lifespan. Cell Syst 2:122-132.
Beck BD, Park SJ, Lee YJ, Roman Y, Hromas RA, Lee SH. 2008. Human Pso4 is a metnase (SETMAR)-binding partner that regulates metnase function in DNA repair. J Biol Chem 283:9023-9030.
Bens M, Sahm A, Groth M, Jahn N, Morhart M, Holtze S, Hildebrandt TB, Platzer M, Szafranski K. 2016. FRAMA: from RNA-seq data to annotated mRNA assemblies. BMC Genomics 17:54.

Biasini M, Bienert S, Waterhouse A, Arnold K, Studer G, Schmidt T, Kiefer F, Gallo Cassarino T, Bertoni M, Bordoli L, et al. 2014. SWISS-MODEL: modelling protein tertiary and quaternary structure using evolutionary information. Nucleic Acids Res 42:W252-258.
Blagosklonny MV. 2008. Aging: ROS or TOR. Cell Cycle 7:3344-3354.
Blagosklonny MV. 2012. Answering the ultimate question "what is the proximal cause of aging?". Aging (Albany NY) 4:861-877.
Bonawitz ND, Chatenay-Lapointe M, Pan Y, Shadel GS. 2007. Reduced TOR signaling extends chronological life span via increased respiration and upregulation of mitochondrial gene expression. Cell Metab 5:265-277.
Bratic A, Larsson NG. 2013. The role of mitochondria in aging. J Clin Invest 123:951-957.
Broer L, Buchman AS, Deelen J, Evans DS, Faul JD, Lunetta KL, Sebastiani P, Smith JA, Smith AV, Tanaka T, et al. 2015. GWAS of longevity in CHARGE consortium confirms APOE and FOXO3 candidacy. J Gerontol A Biol Sci Med Sci 70:110-118.
Buffenstein R. 2008. Negligible senescence in the longest living rodent, the naked mole-rat: insights from a successfully aging species. J Comp Physiol B 178:439-445.
Burda H, Honeycutt RL, Begall S, Locker-Grütjen O, Scharff A. 2000. Are naked and common mole-rats eusocial and if so, why? Behavioral Ecology and Sociobiology 47:293-303.
Chondrogianni N, Petropoulos I, Franceschi C, Friguet B, Gonos ES. 2000. Fibroblast cultures from healthy centenarians have an active proteasome. Exp Gerontol 35:721-728.
Chung HY, Cesari M, Anton S, Marzetti E, Giovannini S, Seo AY, Carter C, Yu BP, Leeuwenburgh C. 2009. Molecular inflammation: underpinnings of aging and age-related diseases. Ageing Res Rev 8:18-30.
Clarke FM, Faulkes CG. 1997. Dominance and queen succession in captive colonies of the eusocial naked mole-rat, Heterocephalus glaber. Proc Biol Sci 264:993-1000.
Cuervo AM, Dice JF. 1996. A receptor for the selective uptake and degradation of proteins by lysosomes. Science 273:501-503.
Dai H, Leeder JS, Cui Y. 2014. A modified generalized Fisher method for combining probabilities from dependent tests. Front Genet 5:32.
Dammann P, Burda H. 2006. Sexual activity and reproduction delay ageing in a mammal. Curr Biol 16:R117-118.
Dammann P, Sell DR, Begall S, Strauch C, Monnier VM. 2012. Advanced glycation end-products as markers of aging and longevity in the long-lived Ansell's mole-rat (Fukomys anselli). J Gerontol A Biol Sci Med Sci 67:573-583.
Dammann P, Sumbera R, Massmann C, Scherag A, Burda H. 2011. Extended longevity of reproductives appears to be common in Fukomys mole-rats (Rodentia, Bathyergidae). PLoS One 6:e18757.
Davies KT, Bennett NC, Tsagkogeorga G, Rossiter SJ, Faulkes CG. 2015. Family Wide Molecular Adaptations to Underground Life in African Mole-Rats Revealed by Phylogenomic Analysis. Mol Biol Evol 32:3089-3107.
de Magalhaes JP, Costa J, Church GM. 2007. An analysis of the relationship between metabolism, developmental schedules, and longevity using phylogenetic independent contrasts. J Gerontol A Biol Sci Med Sci 62:149-160.
Delaney MA, Ward JM, Walsh TF, Chinnadurai SK, Kerns K, Kinsel MJ, Treuting PM. 2016. Initial Case Reports of Cancer in Naked Mole-rats (Heterocephalus glaber). Vet Pathol 53:691-696.
Dillin A, Hsu AL, Arantes-Oliveira N, Lehrer-Graiwer J, Hsin H, Fraser AG, Kamath RS, Ahringer J, Kenyon C. 2002. Rates of behavior and aging specified by mitochondrial function during development. Science 298:2398-2401.
Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15-21.
Duncan EJ, Hyink O, Dearden PK. 2016. Notch signalling mediates reproductive constraint in the adult worker honeybee. Nat Commun 7:12427.
Durlinger AL, Visser JA, Themmen AP. 2002. Regulation of ovarian function: the role of anti-Mullerian hormone. Reproduction 124:601-609.

Estes ML, McAllister AK. 2015. Immune mediators in the brain and peripheral tissues in autism spectrum disorder. Nat Rev Neurosci 16:469-486.
Fan R, Olbricht G, Baker X, Hou C. 2016. Birth mass is the key to understanding the negative correlation between lifespan and body size in dogs. Aging (Albany NY) 8:3209-3222.
Fang X, Nevo E, Han L, Levanon EY, Zhao J, Avivi A, Larkin D, Jiang X, Feranchuk S, Zhu Y, et al. 2014. Genome-wide adaptive complexes to underground stresses in blind mole rats Spalax. Nat Commun 5:3966.
Fang X, Seim I, Huang Z, Gerashchenko MV, Xiong Z, Turanov AA, Zhu Y, Lobanov AV, Fan D, Yim SH, et al. 2014. Adaptations to a subterranean environment and longevity revealed by the analysis of mole rat genomes. Cell Rep 8:1354-1364.
Fletcher W, Yang Z. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol Biol Evol 27:2257-2267.
Flynn JM, Melov S. 2013. SOD2 in mitochondrial dysfunction and neurodegeneration. Free Radic Biol Med 62:4-12.
Frick C, Atanasov AG, Arnold P, Ozols J, Odermatt A. 2004. Appropriate function of 11beta-hydroxysteroid dehydrogenase type 1 in the endoplasmic reticulum lumen is dependent on its N-terminal region sharing similar topological determinants with 50-kDa esterase. J Biol Chem 279:31131-31138.
Fritz-Wolf K, Kehr S, Stumpf M, Rahlfs S, Becker K. 2011. Crystal structure of the human thioredoxin reductase-thioredoxin complex. Nat Commun 2:383.
Fushan AA, Turanov AA, Lee SG, Kim EB, Lobanov AV, Yim SH, Buffenstein R, Lee SR, Chang KT, Rhee H, et al. 2015. Gene expression defines natural changes in mammalian lifespan. Aging Cell 14:352-365.
Gaya-Vidal M, Alba MM. 2014. Uncovering adaptive evolution in the human lineage. BMC Genomics 15:599.
Gorbunova V, Hine C, Tian X, Ablaeva J, Gudkov AV, Nevo E, Seluanov A. 2012. Cancer resistance in the blind mole rat is mediated by concerted necrotic cell death mechanism. Proc Natl Acad Sci U S A 109:19392-19396.
Gorbunova V, Seluanov A, Zhang Z, Gladyshev VN, Vijg J. 2014. Comparative genetics of longevity and cancer: insights from long-lived rodents. Nat Rev Genet 15:531-540.
Groenewoud MJ, Zwartkruis FJ. 2013. Rheb and Rags come together at the lysosome to activate mTORC1. Biochem Soc Trans 41:951-955.
Guruharsha KG, Kankel MW, Artavanis-Tsakonas S. 2012. The Notch signalling system: recent insights into the complexity of a conserved pathway. Nat Rev Genet 13:654-666.
Hall DR, Hadden JM, Leonard GA, Bailey S, Neu M, Winn M, Lindley PF. 2002. The crystal and molecular structures of diferric porcine and rabbit serum transferrins at resolutions of 2.15 and 2.60 A, respectively. Acta Crystallogr D Biol Crystallogr 58:70-80.
Hansen M, Taubert S, Crawford D, Libina N, Lee SJ, Kenyon C. 2007. Lifespan extension by conditions that inhibit translation in Caenorhabditis elegans. Aging Cell 6:95-110.
Hare D, Ayton S, Bush A, Lei P. 2013. A delicate balance: Iron metabolism and diseases of the brain. Front Aging Neurosci 5:34.
Hashemy SI, Holmgren A. 2008. Regulation of the catalytic activity and structure of human thioredoxin 1 via oxidation and S-nitrosylation of cysteine residues. J Biol Chem 283:21890-21898.
Hawes DJ, Brennan J, Dadds MR. 2009. Cortisol, callous-unemotional traits, and pathways to antisocial behavior. Curr Opin Psychiatry 22:357-362.
Helgadottir A, Gretarsdottir S, Thorleifsson G, Hjartarson E, Sigurdsson A, Magnusdottir A, Jonasdottir A, Kristjansson H, Sulem P, Oddsson A, et al. 2016. Variants with large effects on blood lipids and the role of cholesterol and triglycerides in coronary disease. Nat Genet 48:634-639.
Hofmann JW, Zhao X, De Cecco M, Peterson AL, Pagliaroli L, Manivannan J, Hubbard GB, Ikeno Y, Zhang Y, Feng B, et al. 2015. Reduced expression of MYC increases longevity and enhances healthspan. Cell 160:477-488.

Holmes MC, Kotelevtsev Y, Mullins JJ, Seckl JR. 2001. Phenotypic analysis of mice bearing targeted deletions of 11beta-hydroxysteroid dehydrogenases 1 and 2 genes. Mol Cell Endocrinol 171:15-20.
Hughes KA, Reynolds RM. 2005. Evolutionary and mechanistic theories of aging. Annu Rev Entomol 50:421-445.
Hwang J, Nguyen LT, Jeon YH, Lee CY, Kim MH. 2015. Crystal structure of fully oxidized human thioredoxin. Biochem Biophys Res Commun 467:218-222.
Janssens GE, Meinema AC, Gonzalez J, Wolters JC, Schmidt A, Guryev V, Bischoff R, Wit EC, Veenhoff LM, Heinemann M. 2015. Protein biogenesis machinery is a driver of replicative aging in yeast. Elife 4:e08527.
Jarvis JU. 1981. Eusociality in a mammal: cooperative breeding in naked mole-rat colonies. Science 212:571-573.
Jarvis JU, Bennett NC. 1993. Eusociality has evolved independently in two genera of bathyergid mole-rats — but occurs in no other subterranean mammal. Behavioral Ecology and Sociobiology 33:253-260.
Johnson SC, Rabinovitch PS, Kaeberlein M. 2013. mTOR is a key modulator of ageing and age-related disease. Nature 493:338-345.
Jung CH, Ro SH, Cao J, Otto NM, Kim DH. 2010. mTOR regulation of autophagy. FEBS Lett 584:1287-1295.
Kapustin Y, Souvorov A, Tatusova T, Lipman D. 2008. Splign: algorithms for computing spliced alignments with identification of paralogs. Biol Direct 3:20.
Kent WJ. 2002. BLAT--the BLAST-like alignment tool. Genome Res 12:656-664.
Kenyon CJ. 2010. The genetics of ageing. Nature 464:504-512.
Kevei E, Hoppe T. 2014. Ubiquitin sets the timer: impacts on aging and longevity. Nat Struct Mol Biol 21:290-292.
Kim EB, Fang X, Fushan AA, Huang Z, Lobanov AV, Han L, Marino SM, Sun X, Turanov AA, Yang P, et al. 2011. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. Nature 479:223-227.
Kim GH, Kim JE, Rhie SJ, Yoon S. 2015. The Role of Oxidative Stress in Neurodegenerative Diseases. Exp Neurobiol 24:325-340.
Kodavali CV, Watson AM, Prasad KM, Celik C, Mansour H, Yolken RH, Nimgaonkar VL. 2014. HLA associations in schizophrenia: are we re-discovering the wheel? Am J Med Genet B Neuropsychiatr Genet 165B:19-27.
Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six Mammalian genomes. PLoS Genet 4:e1000144.
Lee GY, Haverty PM, Li L, Kljavin NM, Bourgon R, Lee J, Stern H, Modrusan Z, Seshagiri S, Zhang Z, et al. 2014. Comparative oncogenomics identifies PSMB4 and SHMT2 as potential cancer driver genes. Cancer Res 74:3114-3126.
Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754-1760.
Licastro F, Candore G, Lio D, Porcellini E, Colonna-Romano G, Franceschi C, Caruso C. 2005. Innate immunity and inflammation in ageing: a key for understanding age-related diseases. Immun Ageing 2:8.
Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:550.
Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. 2009. GAGE: generally applicable gene set enrichment for pathway analysis. BMC Bioinformatics 10:161.
Macedo MF, de Sousa M. 2008. Transferrin and the transferrin receptor: of magic bullets and other concerns. Inflamm Allergy Drug Targets 7:41-52.
Mahley RW. 2016. Central Nervous System Lipoproteins: ApoE and Regulation of Cholesterol Metabolism. Arterioscler Thromb Vasc Biol 36:1305-1315.
Malkki H. 2016. Neurodevelopmental disorders: Impaired immune system function linked to social behaviour deficits in mice. Nat Rev Neurol 12:431.
Medawar PB. 1952. An unsolved problem of biology. (Printed lecture: University College London).

Mitsui A, Hamuro J, Nakamura H, Kondo N, Hirabayashi Y, Ishizaki-Koizumi S, Hirakawa T, Inoue T, Yodoi J. 2002. Overexpression of human thioredoxin in transgenic mice controls oxidative stress and life span. Antioxid Redox Signal 4:693-696.
Miwa S, Jow H, Baty K, Johnson A, Czapiewski R, Saretzki G, Treumann A, von Zglinicki T. 2014. Low abundance of the matrix arm of complex I in mitochondria predicts longevity in mice. Nat Commun 5:3837.
Nudel R, Simpson NH, Baird G, O'Hare A, Conti-Ramsden G, Bolton PF, Hennessy ER, Consortium SLI, Monaco AP, Knight JC, et al. 2014. Associations of HLA alleles with specific language impairment. J Neurodev Disord 6:1.
Ori A, Toyama BH, Harris MS, Bock T, Iskar M, Bork P, Ingolia NT, Hetzer MW, Beck M. 2015. Integrated Transcriptome and Proteome Analyses Reveal Organ-Specific Proteome Deterioration in Old Rats. Cell Syst 1:224-237.
Perez VI, Cortez LA, Lew CM, Rodriguez M, Webb CR, Van Remmen H, Chaudhuri A, Qi W, Lee S, Bokov A, et al. 2011. Thioredoxin 1 overexpression extends mainly the earlier part of life span in mice. J Gerontol A Biol Sci Med Sci 66:1286-1299.
Pettersen EF, Goddard TD, Huang CC, Couch GS, Greenblatt DM, Meng EC, Ferrin TE. 2004. UCSF Chimera--a visualization system for exploratory research and analysis. J Comput Chem 25:1605-1612.
Pitt JN, Kaeberlein M. 2015. Why is aging conserved and what can we do about it? PLoS Biol 13:e1002131.
Reichwald K, Petzold A, Koch P, Downie BR, Hartmann N, Pietsch S, Baumgart M, Chalopin D, Felder M, Bens M, et al. 2015. Insights into Sex Chromosome Evolution and Aging from the Genome of a Short-Lived Fish. Cell 163:1527-1538.
Roux J, Privman E, Moretti S, Daub JT, Robinson-Rechavi M, Keller L. 2014. Patterns of positive selection in seven ant genomes. Mol Biol Evol 31:1661-1685.
Rubinsztein DC, Marino G, Kroemer G. 2011. Autophagy and aging. Cell 146:682-695.
Saez I, Vilchez D. 2014. The Mechanistic Links Between Proteasome Activity, Aging and Age-related Diseases. Curr Genomics 15:38-51.
Sahm A, Bens M, Platzer M, Cellerino A. 2017. Parallel evolution of genes controlling mitonuclear balance in short-lived annual fishes. Aging Cell.
Sahm A, Bens M, Platzer M, Szafranski K. 2017. PosiGene: automated and easy-to-use pipeline for genome-wide detection of positively selected genes. Nucleic Acids Res.
Sahm A, Platzer M, Cellerino A. 2016. Outgroups and Positive Selection: The Nothobranchius furzeri Case. Trends Genet 32:523-525.
Sanchez-Martin JR, Cardas J, Ahedo L, Fano E, Echebarria A, Azpiroz A. 2001. Social behavior, cortisol, and sIgA levels in preschool children. J Psychosom Res 50:221-227.
Schieke SM, Phillips D, McCoy JP, Jr., Aponte AM, Shen RF, Balaban RS, Finkel T. 2006. The mammalian target of rapamycin (mTOR) pathway regulates mitochondrial oxygen consumption and oxidative capacity. J Biol Chem 281:27643-27652.
Schmidt CM, Bennett NC, Jarvis JU. 2013. The long-lived queen: reproduction and longevity in female eusocial Damaraland mole-rats (Fukomys damarensis). . African Zoology 48:193-196.
Schmidt CM, Blount JD, Bennett NC. 2014. Reproduction is associated with a tissue-dependent reduction of oxidative stress in eusocial female Damaraland mole-rats (Fukomys damarensis). PLoS One 9:e103286.
Seluanov A, Hine C, Azpurua J, Feigenson M, Bozzella M, Mao Z, Catania KC, Gorbunova V. 2009. Hypersensitivity to contact inhibition provides a clue to cancer resistance of naked mole-rat. Proc Natl Acad Sci U S A 106:19352-19357.
Semeiks J, Grishin NV. 2012. A method to find longevity-selected positions in the mammalian proteome. PLoS One 7:e38595.
Simm A, Muller B, Nass N, Hofmann B, Bushnaq H, Silber RE, Bartling B. 2015. Protein glycation - Between tissue aging and protection. Exp Gerontol 68:71-75.
Smorkatcheva AV, Lukhtanov VA. 2014. Evolutionary association between subterranean lifestyle and female sociality in rodents. Mammalian Biology 79:101-109.

Sobrero R, Inostroza-Michael O, Hernandez CE, Ebensperger LA. 2014. Phylogeny modulates the effects of ecological conditions on group living across hystricognath rodents. Animal Behaviour 94:27-34.
Son M, Fu Q, Puttaparthi K, Matthews CM, Elliott JL. 2009. Redox susceptibility of SOD1 mutants is associated with the differential response to CCS over-expression in vivo. Neurobiol Dis 34:155-162.
Stanke M, Schoffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7:62.
Supek F, Bosnjak M, Skunca N, Smuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. PLoS One 6:e21800.
Tacutu R, Craig T, Budovsky A, Wuttke D, Lehmann G, Taranukha D, Costa J, Fraifeld VE, de Magalhaes JP. 2013. Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing. Nucleic Acids Res 41:D1027-1033.
Taylor KR, Milone NA, Rodriguez CE. 2017. Four Cases of Spontaneous Neoplasia in the Naked Mole-Rat (Heterocephalus glaber), A Putative Cancer-Resistant Species. J Gerontol A Biol Sci Med Sci 72:38-43.
Torres AR, Sweeten TL, Cutler A, Bedke BJ, Fillmore M, Stubbs EG, Odell D. 2006. The association and linkage of the HLA-A2 class I allele with autism. Hum Immunol 67:346-351.
Umeda-Kameyama Y, Tsuda M, Ohkura C, Matsuo T, Namba Y, Ohuchi Y, Aigaki T. 2007. Thioredoxin suppresses Parkin-associated endothelin receptor-like receptor-induced neurotoxicity and extends longevity in Drosophila. J Biol Chem 282:11180-11187.
Van Daele PA, Verheyen E, Brunain M, Adriaens D. 2007. Cytochrome b sequence analysis reveals differential molecular evolution in African mole-rats of the chromosomally hyperdiverse genus Fukomys (Bathyergidae, Rodentia) from the Zambezian region. Mol Phylogenet Evol 45:142-157.
Viljakainen L, Evans JD, Hasselmann M, Rueppell O, Tingek S, Pamilo P. 2009. Rapid evolution of immune proteins in social insects. Mol Biol Evol 26:1791-1801.
Vistoli G, De Maddis D, Cipak A, Zarkovic N, Carini M, Aldini G. 2013. Advanced glycoxidation and lipoxidation end products (AGEs and ALEs): an overview of their mechanisms of formation. Free Radic Res 47 Suppl 1:3-27.
Watson WH, Pohl J, Montfort WR, Stuchlik O, Reed MS, Powis G, Jones DP. 2003. Redox potential of human thioredoxin 1 and identification of a second dithiol/disulfide motif. J Biol Chem 278:33408-33415.
Weichsel A, Gasdaska JR, Powis G, Montfort WR. 1996. Crystal structures of reduced, oxidized, and mutated human thioredoxins: evidence for a regulatory homodimer. Structure 4:735-751.
Woodard SH, Fischman BJ, Venkat A, Hudson ME, Varala K, Cameron SA, Clark AG, Robinson GE. 2011. Genes involved in convergent evolution of eusociality in bees. Proc Natl Acad Sci U S A 108:7472-7477.
Wu C, Parrott AM, Fu C, Liu T, Marino SM, Gladyshev VN, Jain MR, Baykal AT, Li Q, Oka S, et al. 2011. Thioredoxin 1-mediated post-translational modifications: reduction, transnitrosylation, denitrosylation, and related proteomics methodologies. Antioxid Redox Signal 15:2565-2604.
Zahn JM, Sonu R, Vogel H, Crane E, Mazan-Mamczarz K, Rabkin R, Davis RW, Becker KG, Owen AB, Kim SK. 2006. Transcriptional profiling of aging in human muscle reveals a common aging signature. PLoS Genet 2:e115.
Zhang C, Cuervo AM. 2008. Restoration of chaperone-mediated autophagy in aging liver improves cellular maintenance and hepatic function. Nat Med 14:959-965.
Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22:2472-2479.

# Diskussion

Die vorliegende Arbeit hat sowohl eine methodische und als auch eine inhaltliche Klammer. Die methodische Klammer ist die genomweite Untersuchung positiver Selektion. Dazu habe ich verschiedene Software-Werkzeuge entwickelt und sie anschließend zum Programm PosiGene zusammengefügt (Manuskript I). PosiGene habe ich dann für meine Beiträge zu anderen Arbeiten eingesetzt (Manuskript II, IV und V). Manuskript III vergleicht die Ergebnisse aus Manuskript II mit denen einer anderen Arbeitsgruppe (Valenzano, et al. 2015). Die Detektion von PSGs auf evolutionären Zweigen, auf denen sehr wahrscheinlich eine Anpassung der Lebensspanne stattgefunden hat, bildet die inhaltliche Klammer der Arbeit (Abbildung 6).

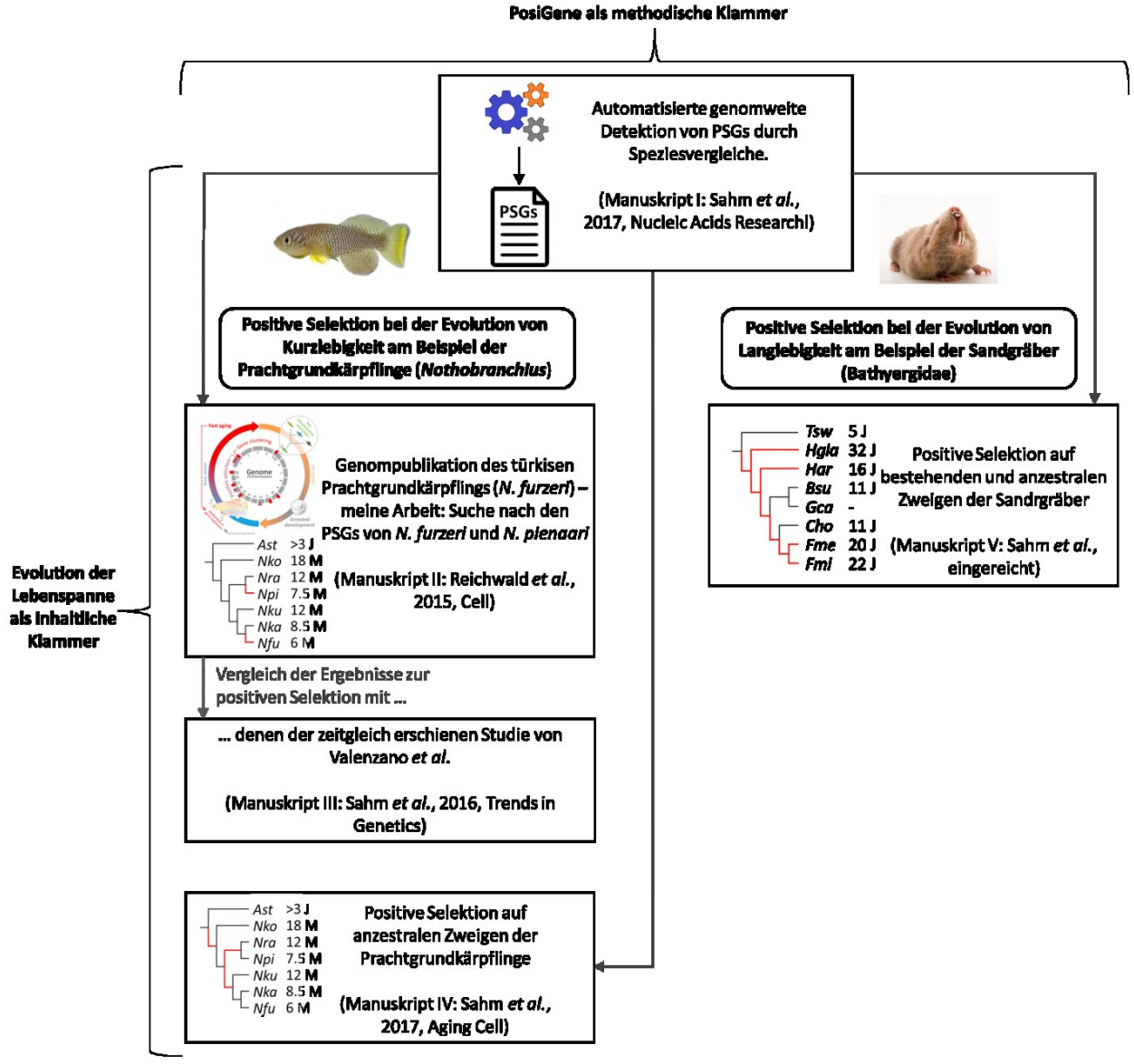



**Abbildung 6.** Zusammenhang der Manuskripte. Das in der Zeitschrift *Nucleic Acids Research* veröffentlichte Programm PosiGene, mit dem PSGs genomweit identifiziert werden können, wurde im Zuge der anderen Arbeiten als ihre jeweilige Hauptmethode angewendet. Drei Arbeiten (links) beschäftigen sich mit positiver Selektion bei Prachtgrundkärpflingen im Kontext der Evolution von Kurzlebigkeit – eine Arbeit (rechts) mit positiver Selektion bei Sandgräbern im Kontext der Evolution von Langlebigkeit. In den Kladogrammen sind Zweige rot markiert, die in den jeweiligen Artikeln nach PSGs abgesucht wurden. Für die Kladogramme links ist die durchschnittlichen Lebensspanne der folgenden Spezies angegeben: *A. striatum*, *N. korthausae*, *N. rachovii*, *N. pienaari*, *N. kuhntae*, *N. kadleci* und *N. furzeri*. Für das Kladogramm rechts sind die maximalen Lebensspannen der folgenden Spezies aufgeführt: *T. swinderianus*, *H. glaber*, *H. argenteocinereus*, *B.suillus*, *G. capensis*, *C. hottentottus*, *F. mechowii* und *F. micklemi*.

## 6. Fortentwicklung und kritische Betrachtung der verwendeten Methode

Die normale Vorgehensweise der auf dem Feld der genomweiten Untersuchung positiver Selektion tätigen Gruppen ist die Entwicklung hausinterner Software, mit deren Hilfe PSGs identifiziert werden, die dann in Publikationen einfließen – letztere werden am Ende des dritten Kapitels und in Manuskript I vielfach zitiert. Die Veröffentlichung der entsprechenden Software wäre in vielen Fällen wahrscheinlich auch nicht sinnvoll, da es sich i.d.R nicht um ein einheitliches Programm zur genomweiten Detektion positiver Selektion, sondern um eine Reihe von einzeln aufzurufenden Programmen handeln dürfte, von denen einige möglicherweise nur im Kontext der Rechnerarchitektur der jeweiligen Arbeitsgruppe funktionieren. Aus diesen Gründen werden die eingesetzten Algorithmen, Filter und externen Programme in den Methodenteilen der entsprechenden Artikel lediglich beschrieben. An diesen Methodenteilen habe ich mich bei der Entwicklung meiner eigenen Programme an verschiedenen Stellen orientiert, so etwa bei der Forderung nach Minimallängen der analysierten Sequenzen (Bakewell, et al. 2007; Davies, et al. 2015) oder bei der Einbindung des Programms GBLOCKS (Talavera and Castresana 2007) für das Filtern problematischer Alignierungsspalten (Seim, et al. 2013; Roux, et al. 2014) – wobei insgesamt festzustellen ist, dass diese Artikel oft heterogen sind, was ihre Methoden anbelangt. Eine andere wichtige methodische Quelle waren daher Empfehlungen der Literatur – so etwa bei der Art der Auswahl der zu alignierenden Isoformen (Villanueva-Canas, et al. 2013) oder bei der Wahl von PRANK (Loytynoja and Goldman 2008) als Alignierungsprogramm für nachfolgende Tests auf positive Selektion (Fletcher and Yang 2010; Privman, et al. 2012). Darauf sowie auf stichprobenartigen, manuellen Inspektionen der Resultate aufbauend habe ich die Programme konzipiert, implementiert und ihre Standardeinstellungen feinjustiert. Ich habe mich dabei, wie im PosiGene-Manuskript beschrieben, vor allem auf die Minimierung der Falsch-Positiv-Rate konzentriert, was auch von der Literatur als das zentrale Problem angesehen wird (Mallick, et al. 2009; Markova-Raina and Petrov 2011).

Die Gründe, die mich dazu bewogen haben, die für diese Arbeit entwickelten Programme entgegen der üblichen Vorgehensweise zu einer Software zusammenzuführen, um diese dann zu publizieren, sind mit dem möglichen Nutzen einer solchen Software für die Wissenschaft verbunden. Zunächst ist zu sagen, dass ich selbst wahrscheinlich nicht mit dem Schreiben der entsprechenden Programme begonnen hätte, wenn es zum damaligen Zeitpunkt bereits eine fertige Software-Lösung gegeben hätte. Das im PosiGene-Manuskript erwähnte Programm POTION (Hongo, et al. 2015), die – wie es im entsprechenden Artikel zurecht heißt – erste Software zur genomweiten Identifizierung von PSGs, wurde erst veröffentlicht als meine Arbeit an PosiGene schon weit vorangeschritten war. Abgesehen davon kann POTION positive Selektion nicht zweigspezifisch detektieren, sondern ausschließlich für den gegebenen Baum insgesamt. Der Umstand, dass POTION damit die wohl am häufigsten verwendete Methode der genomweiten Suche nach PSGs nicht unterstützt, schränkt den praktischen Nutzen deutlich ein (siehe Manuskript I). Ein weiterer Grund für die PosiGene-Publikation war meine Erfahrung, dass es durchaus ein beachtlicher Aufwand war, die nötigen Programme zu implementieren, an dem – bei allem Respekt – wenig bioinformatisch geschulte Personen wahrscheinlich scheitern würden. Daraus folgte für mich, dass eine fertige Software-Lösung nicht nur Bioinformatiker-Kollegen entlasten, sondern auch den Kreis derjenigen, die in der Lage wären genomweit nach PSGs zu suchen erheblich erweitern kann – auf all jene, die einen Kommandozeilenbefehl ausführen können. Des Weiteren würde durch die Publikation einer Software auch ein Qualitätsstandard angeboten, dessen Gewährleistung, wie durch die hohen Falsch-Positiv-Raten in den PSG-Vorhersagen einiger Veröffentlichungen deutlich wird (Markova-Raina and Petrov 2011, Mallick, 2009 #18), ebenfalls nicht trivial ist. Hinzu kommt, dass durch die Veröffentlichung

eines solchen Programms eine einfache Reproduzierbarkeit von Ergebnissen anderer Arbeitsgruppen ermöglicht wird, die gegenwärtig nicht gegeben ist.

Was für die PosiGene-Publikation (Manuskript I) im Vergleich zu anderen nicht veröffentlichten Programmsammlungen zur genomweiten Suche von PSGs hinzukommen musste, sind m.E. vor allem zwei Dinge: Zum Ersten war meine eigene Programmsammlung zu einer Software zu verbinden, die mit nur einem oder wenigen Kommandozeilenbefehlen zu bedienen, entsprechend dokumentiert sowie auf einer großen Zahl von Rechnersystemen ohne komplizierte Installation einsetzbar war. Zum Zweiten – und das war die deutlich größere Herausforderung – musste die Qualität der Software nicht nur durch den Einsatz bewährter bzw. von der Literatur empfohlener Methoden plausibel gemacht, sondern nachgewiesen werden. Das Problem dabei wird durch folgende ironische Frage der Literatur illustriert: „Könnten die echten Loci unter positiver Selektion bitte vortreten?“ ((Biswas and Akey 2006), von mir aus dem Englischen übersetzt). Damit wird auf den Umstand hingewiesen, dass kein Goldstandard existiert, anhand dessen man die Richtig- bzw. Falsch-Positiv-Rate einer PSG-Vorhersagemethode auf einem Realdatensatz zweifelsfrei bestimmen könnte. Aus diesem Grund werden Methoden wie der Zweig-Positionstest zumeist auf Basis simulierter Daten evaluiert. Bei solchen Simulationen werden i.d.R. zunächst eine Reihe von Sequenzen generiert, von denen eine jede entlang der Äste eines gegebenen Baums mutiert wird, sodass jeder Eingangssequenz an der Wurzel des Baums schließlich mehr oder weniger veränderte Sequenzen an den Blättern des Baums gegenüberstehen. Diese Sequenzen an den Blättern des Baums – als das Endprodukt der simulierten Evolution – werden dann aligniert und auf positive Selektion getestet. Da für die Mutation der Sequenzen entlang der Äste definierte Selektionsschemata zum Einsatz kommen, die z.B. positive oder negative Selektion ausdrücken, können richtige von falschen Vorhersagen im Nachhinein präzise unterschieden werden. Der Nachteil dieser Evaluierungsmethode ist, dass das in der Realität auftretende Problemspektrum nur begrenzt abgebildet wird. Das wird bspw. daran deutlich, dass Alignierungsprobleme zwar auf der einen Seite, wie in der Einleitung dargestellt, in realen genomweiten Untersuchungen die Hauptquelle für Falsch-Positive sind (Mallick, et al. 2009; Markova-Raina and Petrov 2011), Simulationen zur Evaluierung des Zweig-Positionstests auf der anderen Seite i.d.R. aber von korrekten Alignierungen ausgingen (z.B. (Zhang, et al. 2005; Nozawa, et al. 2009; Yang and dos Reis 2011; Gharib and Robinson-Rechavi 2013)). D.h. konkret, dass im Rahmen der entsprechenden Evolutionssimulationen keine Insertionen oder Deletionen (InDels) in die Sequenzen eingefügt und die demzufolge gleich langen simulierten Sequenzen an den Blättern des Baums vor Anwendung des Zweigpositionstests durch schlichtes Untereinanderschreiben „aligniert“ wurden. Da auf diese Weise die Codons einer Alignierungsspalte immer in einer Homologiebeziehung zueinander stehen, sind Fehler in der PSG-Vorhersage durch Alignierungsprobleme so *de facto* ausgeschlossen. Aber auch wenn ein Simulationsmodell eingesetzt wird, das nicht nur Nukleotide substituiert, sondern auch InDels einfügt, und die so simulierten Sequenzen vor Anwendung des Zweig-Positionstests einer tatsächlichen und mithin fehlbaren Alignierungsprozedur unterzogen werden, bleiben in der Realität auftretende, mögliche Problemquellen unberücksichtigt, z.B. falsch vorhergesagte Exons, die Existenz von Paralogen und Pseudogenen, Assemblierungsartefakte wie chimäre Sequenzen etc. Es existiert keine Simulationsmethode, die derartige, in Realdaten auftretende Herausforderungen für die Untersuchung positiver Selektion abzubilden vermag. Der Aufwand, der für die Entwicklung einer solchen Methode betrieben werden müsste, würde den der Entwicklung von PosiGene wahrscheinlich übersteigen. Hinzu kommt, dass auch im Rahmen verfügbarer Simulationsmethoden eine Reihe von Annahmen gemacht werden müssen, z.B. über die Längenverteilung der einzufügenden InDels, über die Häufigkeit ihres Auftretens oder die Stärke und die Verteilung des simulierten Selektionsdrucks. Dabei besteht zumindest die Gefahr, dass die Realität durch die Festlegung solcher Parameter nur unzureichend

widergespiegelt wird. Eine andere Möglichkeit der Validierung wäre PSGs auf einem realen Datensatz zu bestimmen und die Ergebnisse mit denen bereits existierender Studien zu vergleichen. Zwar existieren, wie bereits mehrfach erwähnt, viele Artikel, die positive Selektion untersuchen – allerdings in den meisten Fällen nicht mehr als einer pro untersuchtem evolutionären Zweig. Da sich die Güte anderer PSG-Vorhersagemethoden ebenso wenig einschätzen lässt, wie die der eigenen Methode, wäre ein Vergleich mit den Ergebnissen nur einer anderen Studie aber wenig aussagekräftig.

Um der im vorhergehenden Absatz beschriebenen Problemlage angemessen zu begegnen, habe ich eine komplementäre Validierungsstrategie entwickelt: Zum einen habe ich die Richtig- und die Falsch-Positiv-Rate der Vorhersagen auf Basis von Simulationen bestimmt, die auch InDels modellieren, sodass ein Potenzial für Alignierungsprobleme grundsätzlich gegeben war, und mich bei der Festlegung der Parameter der Simulation soweit als möglich auf empirische Werte aus der Literatur gestützt. Beispielsweise habe ich das Verhältnis von Substitutionen zu InDels auf 43:1 festgelegt, weil dies dem in den kodierenden Sequenzen von Primaten gefundenen Verhältnis entspricht (Chen, et al. 2009). Zum anderen habe ich PosiGene auf Realdaten validiert, indem ich die PSG-Vorhersagen von PosiGene mit denen mehrerer anderer genomweiter Untersuchungen positiver Selektion verglichen habe. Die zugrundeliegende Annahme ist, dass PSGs, die von mehreren Studien unabhängig voneinander gefunden wurden, verlässlicher sind als solche, die nur von einer Studie vorhergesagt werden. Als Untersuchungsgegenstand bot sich der Mensch an, da der entsprechende evolutionäre Zweig einer der wenigen ist, die von verschiedenen Arbeitsgruppen auf Basis von Speziesvergleichen genomweit nach PSGs abgesucht wurden.

Im Ergebnis wurde durch die Validierung auf den simulierten Daten demonstriert, dass PosiGene PSGs mit einer sehr geringen Falsch-Positiv-Rate (0,3-0,4%) vorhersagt. Dies entspricht einem Bruchteil der aus der Literatur bekannten Falsch-Positiv-Rate auf ungefilterten Daten (2,1-13%, (Fletcher and Yang 2010)). PosiGene‘s strikte Filtermechanismen unterdrücken also Falsch-Positive effektiv. Zwar war dies, wie zuvor beschrieben, das Hauptziel bei der Implementierung gewesen – verlässliche statt vieler PSG-Vorhersagen – dennoch stellte sich die Frage, inwieweit die niedrige Falsch-Positiv-Rate durch eine niedrige Richtig-Positiv-Rate erkauft war. Tatsächlich liegt PosiGene’s Richtig-Positiv-Rate auf Basis der Simulation (5,4-30,7%) im oberen Bereich dessen, was mit dem Zweig-Positionstest technisch maximal möglich ist (1,4-33,1%, (Fletcher and Yang 2010)). Die Filter-Mechanismen entfernen also nur wenige echte Signale positiver Selektion. Mit Blick auf die Anwendungen in dieser Arbeit ist trotzdem zu berücksichtigen, dass die von PosiGene detektieren PSGs auf der einen Seite zwar wahrscheinlich korrekt sind, aber auf deren anderen Seite nur einen Ausschnitt der tatsächlich vorhandenen positiven Selektion auf kodierenden Sequenzen repräsentieren. Die Ergebnisse der Validierung auf den Realdaten zeigen, dass zwei Drittel der von PosiGene auf dem Human-Zweig identifizierten PSGs von mindestens einer anderen, hochrangigen Studie ebenfalls vorhergesagt werden – die Hälfte der von PosiGene ermittelten PSGs werden sogar von zwei anderen Studien bestätigt. Auf Basis dieser Maßstäbe hat PosiGene im Vergleich von sieben genomweiten Untersuchungen positiver Selektion die jeweils niedrigste Falsch-Positiv-Rate (wenn also PSGs, die von nur einer Untersuchung identifiziert wurden, als Falsch-Positive definiert werden).

PosiGene ist dazu gedacht mit der genomweiten Suche nach PSGs eine bereits beliebte und erfolgreiche Methode weiter zu verbessern, indem einerseits Schwellen für ihren Einsatz – wie das dazu nötige Wissen und der Aufwand an Arbeitszeit – weitgehend beseitigt werden und andererseits eine hohe Qualität der Ergebnisse garantiert wird. Das kommt vor allem jenem Ziel zu Gute, das, wie im dritten Kapitel dargelegt, das Ziel der meisten Arbeiten war, die diese Methode angewendet haben: PSGs mit

Änderungen von Phänotypen in Verbindung zu bringen, die auf bestimmten evolutionären Zweigen stattgefunden haben. Auf diesem Wege konnten durch die bisherigen genomweiten Untersuchungen positiver Selektion, wie in der Einleitung und im PosiGene-Manuskript beschrieben, Einsichten in evolutionäre Anpassungsprozesse und in die genetische Basis speziesspezifischer phänotypischer Eigenschaften gewonnen werden. Ich erhoffe mir, dass PosiGene zu weiteren derartigen Erkenntnissen beitragen kann, indem das Programm bspw. wie in dieser Arbeit als methodische Grundlage für Studien mit dem Fokus auf positive Selektion dient oder neben anderen Analysen im Rahmen von Genomprojekten eingesetzt wird.

Mit Blick auf das Ziel PSGs mit evolutionären Veränderungen bestimmter Phänotypen zu assoziieren, sollen einige wichtige Einschränkungen benannt werden. Erstens, handelt es sich bei der Identifizierung eines Gens als PSG um eine auf einem Modell beruhende Vorhersage. Das Modell und seine Annahmen wurden im dritten Kapitel erörtert, ebenso wie der Umstand, dass viele Annahmen schwer zu überprüfen sind und in der Realität nicht immer zutreffen müssen. Dies möchte ich anhand von drei Beispielen illustrieren. Eine der grundlegenden Annahmen aller $d_N/d_S$-basierten Methoden, einschließlich des Zweig-Positionstest, ist bspw., dass synonyme Substitutionen annähernd neutral selektiert werden, weil die Änderung eines Nukleotids einer kodierenden Sequenz, die nicht gleichzeitig eine Aminosäureänderung im kodierten Protein bewirkt, keine funktionelle Relevanz haben könne. Auf Basis dieser Annahme wird, wie in der Einleitung erläutert, die Rate der synonymen Substitutionen $d_S$ als Maßstab verwendet, um festzustellen ob die Rate der nicht-synonymen Substitutionen $d_N$ in signifikanter Weise beschleunigt ist, was ggf. als starkes Indiz für positive Selektion zu werten ist. Obwohl die Annahme der funktionellen Irrelevanz synonymer Substitutionen generell plausibel und weit verbreitet ist, gibt es Hinweise darauf, dass zumindest einige spezielle Fälle nicht-neutraler Selektion synonymer Substitutionen existieren. So wurde Selektion synonymer Substitutionen z.B. mit der Effizienz der Transkription (Xia 1996) und des Spleiß-Vorgangs (Parmley, et al. 2006), mit der Effizienz (Carlini and Stephan 2003) und Genauigkeit (Zhou, et al. 2012) der Translation sowie mit der Stabilität der Sekundarstrukturen von DNA (Vinogradov 2003) und mRNA (Stoletzki 2008) in Verbindung gebracht. Eine weitere Annahme bei der genomweiten Anwendung des Zweig-Positionstests ist, dass $d_S$ zwar zwischen verschiedenen Genen variieren kann, aber über eine einzelne kodierende Sequenz hinweg konstant ist. Es wird also einerseits davon ausgegangen, dass verschiedene Gene in unterschiedlichem Maße vom evolutionären Hintergrundrauschen durch neutrale selektierte Substitutionen betroffen sein können und die jeweiligen $d_N$-Werte mit Hinblick auf die Detektion positiver Selektion auch entsprechend unterschiedlich bewertet werden müssen Als Grund für das Variieren von $d_S$ zwischen den Genen wird die unterschiedliche mechanistische Anfälligkeit der die kodierenden Sequenzen beherbergenden Genombereiche gegenüber Mutationen angesehen – z.B. durch den GC-Gehalt oder den Replikationszeitpunkt des Bereichs in der S-Phase der Zelleteilung (Hodgkinson and Eyre-Walker 2011). Es existieren Hinweise darauf, dass sich diese Anfälligkeit gegenüber Mutationen innerhalb 100 kb großer Genombereiche kaum unterscheidet, dass bei der Betrachtung größerer Bereiche die Variation der Mutationsrate aber exponentiell zunimmt, bis bei ca. 10-15 Mb großen Bereichen keine entsprechende Autokorrelation mehr feststellbar ist (Gaffney and Keightley 2005). Die längsten humanen, proteinkodierenden Gene sind aufgrund vieler langer Introns über 2 Mb groß – z.B. *DMD* (Sakharkar, et al. 2004); allerdings sind auch 99% bzw. 84% der humanen, proteinkodierenden Gene kürzer als 500 bzw. 100 kb (eigene Berechnung basierend auf den Daten von Ensembl BioMart, (Kinsella, et al. 2011)). Davon ausgehend wäre es also denkbar, dass $d_S$ entgegen den Modellannahmen in einigen Fällen über die kodierende Sequenz hinweg Schwankungen unterworfen ist. Dies könnte sowohl zu Falsch-Negativen als auch Falsch-Positiven Vorhersagen führen. Die letzte Annahme des Zweig-Positionstests, die ich hier

besprechen will, besteht darin, dass sein zugrundeliegendes Modell zwar erlaubt, dass mehrere Nukleotide eines Codons auf einem Ast substituiert werden – aber nur nacheinander ($\Delta t > 0$) und nicht gleichzeitig ($\Delta t = 0$). Diese Annahme führt u.a. dazu, dass bei geringen $d_N$- und $d_S$-Werten häufig schon ein einziges Codon mit Austauschen an zwei oder drei seiner Positionen ausreicht, damit positive Selektion detektiert wird (sogenannter Suzuki-Effekt, Nozawa, et al. 2009). Der Grund dafür ist, dass es unter der Bedingung geringer Sequenzdivergenz als sehr unwahrscheinlich gelten darf, dass mehrere nicht-synonyme Substitutionen zufällig nacheinander das gleiche Codon betreffen (Yang and dos Reis 2011). Diese Schlussfolgerung gilt aber nicht, falls doch mehrere Nukleotide auf einmal ausgetauscht werden können. Wie häufig die soeben diskutierten Annahmen tatsächlich verletzt werden und inwieweit dies die Vorhersagen des Zweig-Positionstests beeinträchtigt ist weitgehend unerforscht. Unabhängig davon ist zu berücksichtigen, dass eine PSG-Vorhersage auch unter der Voraussetzung, dass die Annahmen zutreffen, eine durch den p-Wert ausgedrückte α-Fehlerwahrscheinlichkeit besitzt. Der endgültige Nachweis für die funktionelle Relevanz eines PSGs bzw. seiner positiv selektierten Positionen kann daher nur im Labor erfolgen.

Selbst wenn die Sequenzänderungen von PSGs funktionell relevant sind, müssen sie nicht zwangsläufig mit den phänotypischen Veränderungen verbunden sein, die für diejenigen, die die Methode anwenden, am interessantesten erscheinen. Es ist davon auszugehen, dass auf vielen evolutionären Zweigen mehr als nur eine phänotypische Eigenschaft angepasst wird. So konzentriert sich diese Arbeit auf evolutionäre Sequenzänderungen, die für die Alternsrate relevant sein könnten, dennoch haben z.B. mit Blick auf die untersuchten Sandgräber-Zweige sehr wahrscheinlich Anpassungen hinsichtlich der ebenfalls auf diesen Zweigen entwickelten eusozialen und Untergrundlebensweise stattgefunden. Da die Evolution der Eusozialität bei Sandgräbern ein in der Literatur kontrovers diskutiertes Thema ist (Jarvis and Bennett 1993; Burda, et al. 2000), haben wir neben dem Thema Alterung im entsprechenden Manuskript die PSGs auch in dieser Richtung interpretiert. Der Evolution der unterirdischen Lebensweise haben wir in diesem Artikel hingegen keine Aufmerksamkeit geschenkt. Ein PSG auf dem Silbermullzweig, das darin involviert sein könnte, ist z.B. *HBB*, das für eine Hämoglobinuntereinheit kodiert und somit essentiell für den Transport von Sauerstoff ist, dessen Konzentration in den unterirdischen Bauten der Sandgräber niedrig ist (Bennett and Faulkes 2000). Neben solchen phänotypischen Veränderungen wie sie gerade beispielhaft angesprochen wurden, ist auch denkbar, dass ermittelte PSGs mit subtileren evolutionären Anpassungen der Spezies assoziiert sind, derer wir uns schlicht nicht bewusst sind.

Es ist wichtig sich klarzumachen, dass die Interpretation der PSGs in unseren wie in den anderen auf diesem Feld tätigen Arbeitsgruppen im Wesentlichen im Lichte des bereits zu diesen Genen sowie zu den betrachteten biologischen Prozessen vorhandenen Wissens durchgeführt werden. Das ist offensichtlich, wenn in unseren Manuskripten bspw. zu einzelnen PSGs alternsrelevante Erkenntnisse aus der Literatur – wie die Assoziation mit bestimmten Krankheiten – zusammengefasst werden. Das Gleiche trifft aber auch zu, wenn z.B. statistisch signifikante Anreicherungen der PSGs in alternsrelevanten funktionellen, biologischen Kategorien wie z.B. „Translation“ oder „Mitochondriale Biogenese“ berichtet werden, da die Voraussetzung für solche Tests eine Datenbank ist, in der entsprechende Erkenntnisse aus der Forschung abgelegt wurden. Das heißt zum einen, dass es wahrscheinlich ist, dass in den rohen, oft Dutzende oder Hunderte Einträge umfassenden PSG-Listen, die aufgrund dieser Größe zumeist nur in den Anhängen der jeweiligen Arbeiten publiziert werden, noch viele unerkannte genetische Grundlagen evolutionärer Anpassungen schlummern. Auf der anderen Seite ist es auch möglich, dass ein Gen obwohl es auf einem Zweig selektiert wurde, auf dem evolutionäre Veränderungen einer phänotypischen Eigenschaft stattgefunden haben, die zu dem passt, was man bislang über die Funktion des Gens weiß, es dennoch in Zusammenhang mit der Veränderung einer anderen Eigenschaft positiv selektiert wurde. Auch an dieser

Stelle könnten Experimente *in vivo* oder *in vitro* dabei helfen, die *in silico* Ergebnisse bzw. daraus abgeleitete biologische Hypothesen zu verifizieren.

Daher ist es m.E. ein nicht unwesentliches Problem für das Feld der Untersuchung positiver Selektion, dass bislang kaum Beispiele existieren, in denen einzelne PSGs und ihre positiv selektierter Positionen im Labor gezielt auf ihre biologische Funktion hin untersucht wurden (Yokoyama 2013). Eines dieser wenigen Beispiele ist *TRIM5α*, das für einen Restriktionsfaktor kodiert, der Infektionen durch Retroviren entgegenwirkt. Das Gen ist positiv selektiert, sowohl auf anzestralen Primatenzweigen ebenso wie im Menschen als auch in einigen heute lebenden Affenarten. Es wurde gezeigt, dass ein 13 Aminosäuren langes Segment, das die mit Abstand höchste Konzentration positiv selektierter Positionen des PSGs enthält, verantwortlich für die meisten speziesspezifischen antiviralen Restriktionsaktivitäten ist. Insbesondere vermitteln Austausche an den positiv selektierten Positionen in diesem Segment die Restriktionskapazität des entsprechenden Rhesusaffenproteins gegenüber HIV und tragen somit sehr wahrscheinlich entscheidend zur Immunität der Rhesusaffen gegen HIV bei (Sawyer, et al. 2005). Auf der anderen Seite wurde gezeigt, dass Austausche an einer Reihe positiv selektierter Positionen in den Rhodopsinen – also in Sehpigmenten – verschiedener Fischspezies keinen Einfluss auf das Lichtabsorptionsmaximum der entsprechenden Genprodukte haben (Yokoyama, et al. 2008). Es mag allerdings sein, dass die Effektgröße unter der damaligen Nachweisbarkeitsgrenze lag oder dass die Mutationen die Fitness in anderer Weise beeinflussen als durch die Veränderung des Absorptionsmaximums (Yang and dos Reis 2011). Weiter wurde durch gerichtete Mutagenese an den auf einem anzestralen Säugerzweig positiv selektierten Positionen des immunrelevanten Enzyms MPO (Myeloperoxidase) nachgewiesen, dass die entsprechenden Aminosäureaustausche essentiell sind für die Fähigkeit des Enzyms das starke Mikrobizid Hypochlorige Säure zu produzieren (Loughran, et al. 2012). Dass es nur wenige derartige experimentelle Folgestudien gibt, mag mehrere Gründe haben. Zum einen ist sicher der Aufwand solcher Studien zu berücksichtigen. Gerade was Experimente *in vivo* anbelangt, war das Erzeugen von genetisch veränderten Organismen mit gerichteten Punktmutationen bis zur Erfindung der CRISPR/Cas-Methode vor wenigen Jahren noch sehr zeit- und kostenintensiv (Wang, et al. 2013). Zum anderen existierte m.E. und existiert z.T. noch immer ein gewisser personeller und inhaltlicher Graben zwischen den Arbeitsgruppen, die bioinformatische Vorhersagen wie die Identifizierung von PSGs tätigen und den Gruppen, die die Expertise für entsprechende Laborexperimente haben, der nur allmählich überbrückt wird. Hinzu kommt, dass die genaue Kenntnis der positiv selektierten Positionen im jeweiligen Sequenzkontext sowie der Aminosäuren der verschiedenen Spezies an diesen Positionen, die Voraussetzung für weiterführende Experimente an einzelnen PSGs ist. Die meisten genomweiten Untersuchungen positiver Selektion haben diese Informationen aus ihren Ergebnissen aber nicht in entsprechenden Alignierungsvisualisierungen zusammengeführt oder diese zumindest nicht veröffentlicht. Unter anderem aus diesem Grund erstellt PosiGene automatisch solche Alignierungsvisualisierungen für jedes Resultat, sodass biologische Experten – nach Durchsicht der Visualisierungen der ihrer Meinung nach interessanten PSGs – Hypothesen formulieren und basierend darauf ggf. Experimente planen können.

## 7. Positive Selektion bei kurzlebigen Prachtgrundkärpflingen

Die erste Anwendung der Programme, die später zu PosiGene zusammengefasst wurden, erfolgte im Rahmen der Genompublikation des Türkisen Prachtgrundkärpflings (*N. furzeri*, Manuskript II). Wie u.a. aus der 31 Namen umfassenden Autorenliste hervorgeht, war dies ein – jedenfalls verglichen mit den anderen in diese Arbeit eingebundenen Manuskripte – sehr umfangreiches Projekt (vgl. Übersicht der Manuskripte). Wie die meisten anderen Koautoren, hatte ich lediglich einen kleinen Ausschnitt des

Gesamtprojekts zu verantworten. Mit der Veröffentlichung der Genomsequenz wurde vor allem das Ziel verfolgt, eine der entscheidenden Voraussetzungen für die Etablierung des Türkisen Prachtgrundkärpflings als Modellorganismus der Alternsforschung zu schaffen. Bspw. ist die Verfügbarkeit von Genom- bzw. Transkriptomsequenzen die Voraussetzung für viele bioinformatorische Analysen, wie z.B. für das Bestimmen differentiell exprimierter Gene oder für die Anwendung von Methoden aus dem Bereich der komparativen Genomik. Eine Reihe derartiger Analysen wurde in Manuskript II eingesetzt, um neben der Bereitstellung des Genoms als Ressource zusätzlich direkte Beiträge einerseits zur Alternsforschung zu leisten und andererseits eine Theorie zum genetischen Geschlechtsbestimmungsmechanismus des Türkisen Prachtgrundkärpflings zu entwickeln.

Mein Hauptziel im Zusammenhang mit Manuskript II war es nach PSGs auf dem Zweig des Türkisen Prachtgrundkärpflings zu fahnden, um so Gene zu identifizieren, die relevant für die evolutionäre Anpassung seiner phänotypischen Eigenschaften, insbesondere der Lebensspanne, sein könnten. Das Gleiche habe ich außerdem für den Zweig des *N. pienaari* (Trivialname existiert nicht) durchgeführt, dessen durchschnittliche Lebensspanne mit 7,5 Monaten fast so kurz ist wie des Türkisen Prachtgrundkärpflings (6 Monate). Zur Detektion positiver Selektion auf diesen beiden Zweigen wurden die proteinkodierenden Sequenzen (CDSs) von *N. furzeri* und *N. pienaari* mit denen Arten der Gattung *Nothobranchius* verglichen, die eine höhere Lebenserwartung aufweisen. Hinzu kamen Sequenzen des gestreiften Prachtkärpflings (*Aphyosemion striatum*) als Vertreter der nächsten nicht-einjährigen, d.h. nicht dem für die *Nothobranchius*-Arten charakteristischen jährlichen Sterberhythmus unterworfenen, Verwandten der Gattung (durchschnittliche Lebensspanne > 3 Jahre, Manuskript II: Abbildung 7). Was die Generierung der Sequenzdaten anbelangt, war ich auf die Vorarbeit von Kollegen angewiesen, die Gehirnproben aller sieben Fischarten asservierten, RNA isolierten, RNA-seq durchführten, die Sequenzdaten zu Transkripten assemblierten und auf diesen jeweils die CDS identifizierten. Diese Datensätze habe ich benutzt und mit PosiGene analysiert

Im Ergebnis wurden auf dem Zweig des Türkisen Prachtgrundkärpflings unter 11.748 Genen sieben unter positiver Selektion detektiert und auf dem Zweig des *N. pienaari* eines unter 5.576. Für die Alternsrelevanz zumindest eines Teils dieser Gene, spricht neben dem Umstand der positiven Selektion auf evolutionären Zweigen, auf denen sehr wahrscheinlich die Lebensspanne verkürzt wurde, zum einen, dass sich die Expressionshöhe von fünf dieser Gene in mindestens einem von drei Geweben des Türkisen Prachtgrundkärpflings während des Alterns signifikant verändert (Gehirn, Leber und Haut; 5 Wochen gegen 39 Wochen, Tabelle 1). Zum anderen sind zwei dieser Gene im Rahmen der Alternsforschung bereits bekannt: *ID3* ist eine Schlüsselkomponente des TGF-β Signalwegs, der Entzündungen reguliert und mit mehreren alternsbedingten Krankheiten wie chronischem Leber- und Nierenversagen, neurodegenerativen Erkrankungen, Arthrose und Krebs assoziiert ist (Krieglstein, et al. 2012). Mit Hilfe biologischer Experten wurde zudem konkret ein Abschnitt des Gens hervorgehoben, an dem auf eine radikale Substitution an einer positiv selektierten Position eine 2 Aminosäuren lange Deletion folgt (siehe Abbildung 4D in Manuskript II). IKBIP befördert die Apoptose (Hofer-Warbinek, et al. 2004) und damit einen alternsrelevanten Prozess (Shen and Tower 2009), der im Alter beim Türkisen Prachtgrundkärpfling verstärkt induziert wird (Ng'oma, et al. 2014).

**Tabelle 1.** Positiv selektierte Gene in Manuskript II.

| Gensymbol | Differentiell exprimiert während des Alterns im Türkisen Prachtgrundkärpfling* |
|---|---|
| **Untersuchter Zweig: Türkiser Prachtgrundkärpfling** | |
| *IKBIP* | ↓ Haut |
| *BX511257* | ↓ Gehirn |
| *ANKZF1* | - |
| *DAGLB* | ↑ Haut |
| *ID3* | ↑ Gehirn & Haut |
| *ANO4* | ↑ Gehirn & Haut |
| *ZDHHC7* | - |
| **Untersuchter Zweig: *Nothobranchius pienaari*** | |
| *CLCC1* | - |

* 5 Wochen gegen 39 Wochen alte Türkise Prachtgrundkärpflinge; Gewebe: Gehirn, Leber, Haut

Abseits von der Kernthematik dieser Arbeit habe ich für Manuskript II einen Beitrag zur Identifikation des wahrscheinlich geschlechtsbestimmenden Gens *GDF6* geleistet. Dazu habe ich die männliche und weibliche Sequenz der 339 Gene in der zuvor von Kollegen identifizierten geschlechtsbestimmenden Region verglichen und festgestellt, dass *GDF6* das stärkste lokale Signal positiver Selektion enthält.

Zeitgleich mit Manuskript II wurde in derselben Ausgabe des Journals *Cell* eine weitere Genompublikation zum Türkisen Prachtgrundkärpfling von einer Arbeitsgruppe aus Stanford veröffentlicht (Valenzano, et al. 2015). Keine der beiden Arbeitsgruppen war über die konkreten Inhalte der jeweils anderen Arbeit informiert. Nach der Veröffentlichung zeigte sich, dass die Gruppe aus Stanford ebenfalls positive Selektion beim Türkisen Prachtgrundkärpfling untersucht und die Ergebnisse aus dem Blickwinkel der Alternsforschung interpretiert hatte. Während wir in Manuskript II, wie dargestellt, sieben PSGs detektiert haben, wurden von Valenzano *et al.* allerdings mehr als 500 PSGs identifiziert. Die Frage, die ich nach der Lektüre des Manuskripts der Kollegen aus Stanford beantworten wollte, war: Wie ist dieser große Unterschied zu erklären? Aus dieser Fragestellung heraus entstand Manuskript III.

Valenzano *et al.* hatten neben den CDSs des Türkisen Prachtgrundkärpflings, die sie im Rahmen ihres Genomprojekts, assembliert und annotiert hatten, ausschließlich öffentliche Daten verwendet, was die Vergleichsspezies angeht. Das führte, aufgrund des Mangels an verfügbaren, bereits sequenzierten Fischgenomen dazu, dass die am engsten mit dem Türkisen Prachtgrundkärpfling verwandte Spezies in der Analyse der Gruppe aus Stanford der Spiegelkärpfling (*Xiphophorus maculatus*) war. Der letzte gemeinsame Vorfahr des Spiegelkärpflings und des Türkisen Prachtgrundkärpflings lebte vor ca. 50-70 Millionen Jahren (Near, et al. 2012). Damit ist der der Spiegelkärpfling evolutionär weiter von den *Nothobranchius*-Arten entfernt als der in unserer Analyse als Außengruppe eingesetzte gestreifte Prachtkärpfling (*Aphyosemion striatum*).

Der von uns untersuchte Zweig des Prachtgrundkärpflings (Abbildung 7, Zweig 6) ist mithin also nur ein Teilstück des – einen viel größeren Zeitraum überspannenden – Zweigs den Valenzano *et al.* untersucht hatten (Zweig 1-6). Daher konnte ich davon ausgehen, dass auch nur ein Teil der von Valenzano *et al.* identifizierten positiv selektierten Positionen tatsächlich spezifisch für den Türkisen Prachtgrundkärpfling (also Zweig 6) sein würden. Da mir die Sequenzen jener näher mit dem Türkisen Prachtgrundkärpfling verwandten Arten aus Manuskript II zur Verfügung standen, die die Kollegen aus Stanford noch nicht in ihre Analyse einbeziehen konnten, war ich in der Lage dies zu überprüfen. Die höhere phylogenetische Auflösung erlaubte mir, die Substitutionen an den von Valenzano *et al.* vorhegesagten positiv selektierten Positionen hinsichtlich ihres Auftretens im phylogenetischen Baum sehr viel genauer zu datieren, als es

die Kollegen aus Stanford mit ihren Daten konnten. Dazu inspizierte ich vor allem jene PSGs, die Valenzano *et al.*im Haupttext erwähnt und dementsprechend als möglicherweise alternsrelevant interpretiert hatten (für ein konkretes Beispiel siehe Manuskript III: Abbildung 2).

Die Frage der Alternsrelevanz ist allerdings zu trennen von der Frage, ob die positiv selektierten Positionen spezifisch für den Türkisen Prachtgrundkärpfling sind. Vielmehr ist es, wie in der Einleitung erläutert und aus dem phylogenetischen Baum im Zusammenspiel mit den Lebensspannen ersichtlich, sehr wahrscheinlich, dass es ebenfalls auf anzestralen Zweigen der Prachtgrundkärpflinge zu Verkürzungen der Lebensspanne gekommen ist (Abbildung 7, Zweig 2-5). Substitutionen, die allerdings schon auf dem Zweig des letzten gemeinsamen Vorfahren des Spiegelkärpflings und des gestreiften Prachtkärpflings stattgefunden haben (Abbildung 7, Zweig 1), sind hingegen mit sehr großer Wahrscheinlichkeit nicht alternsrelevant. Schließlich sind beide Spezies mit einer durchschnittlichen Lebensspanne von mindestens 30 Monaten für ihre Größe nicht kurzlebig (Margolis-Nunno, et al. 1986; Reichwald, et al. 2015) und die einjährige Lebensweise sowie die kürzeren Lebensspannen ist nach allem, was man weiß, erst in der Klade Prachtgrundkärpflinge entstanden (Furness, et al. 2015).

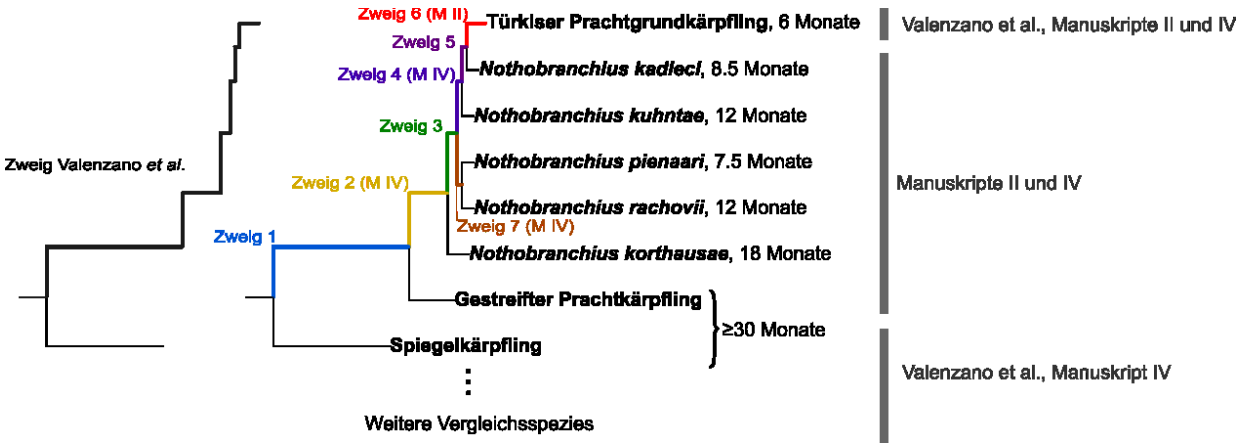



**Abbildung 7.** Analysierte Zweige in Manuskript II, IV und Valenzano *et al*. Die Zweigbezeichnungen entsprechen der Notation aus Manuskript III, das die Ergebnisse von Manuskript II und Valenzano *et al*. (Valenzano, et al. 2015) vergleicht. In Klammern sind ggf. die Manuskripte angegeben, in denen die Zweige jeweils nach PSGs abgesucht wurden. Die in Manuskript IV untersuchten Zweige 2, 4 und 7 heißen in der Notation von Manuskript IV N, FKK bzw. PR. Links ist zur Veranschaulichung der Zweig dargestellt, der in Valenzano *et al*. nach PSGs abgesucht wurde, da dort der Spiegelkärpfling als nächstverwandte Vergleichsspezies zum Türkisen Prachtgrundkärpfling eingesetzt wurde. Die Monatsangaben rechts neben den Speziesnamen geben die durchschnittliche Lebensspanne der jeweiligen Art an (Margolis-Nunno, et al. 1986; Reichwald, et al. 2015). Rechts außen ist dargestellt, welche Arten in welchen Manuskripten in die jeweilige Analyse einbezogen worden sind.

Im Ergebnis habe ich festgestellt, dass nur ein kleiner Bruchteil der Substitutionen an den positiv selektierten Positionen der von Valenzano *et al.* als potentiell alternsrelevant diskutierten PSGs im Türkisen Prachtgrundkärpfling stattgefunden hat (2 von 46, Tabelle 2). Anhand dieses Stichprobenresultats kann der große Unterschied in der PSG-Zahl der beiden Analysen weitgehend damit erklärt werden, dass die untersuchten Zweige nur dem Namen nach identisch sind, tatsächlich aber verschiedene phylogenetische Entitäten repräsentieren. Mit Blick auf die zweite Fragestellung – also der möglichen Alternsrelevanz – ist festzustellen, dass die deutliche Mehrzahl der Substitutionen an den positiv selektierten Positionen nicht nur in der Summe, sondern auch in jedem einzelnen der betrachteten PSGs auf Zweig 1 stattgefunden haben (Tabelle 2). Der Großteil der Substitutionen repräsentiert also Anpassungen, die vor der evolutionären Verkürzung der Lebensspannen stattgefunden haben. Die meisten dieser Signale positiver Selektion sind daher im Widerspruch zur Interpretation durch Valenzano *et al.* mit sehr großer Wahrscheinlichkeit nicht alternsrelevant.

**Tabelle 2.** Datierung der Substitutionen an den positiv selektierten Positionen der in Valenzano *et al.* als potenziell alternsrelevant diskutierten PSGs.

| Gensymbol | Zweig 1 | Zweig 2 | Zweig 3 | Zweig 4 | Zweig 6 | Alignierungs-/Sequenzprobleme | Summe |
|---|---|---|---|---|---|---|---|
| *LMNA* | 2 | 0 | 0 | 0 | 0 | 0 | 2 |
| *BAX* | 2 | 1 | 0 | 0 | 0 | 0 | 3 |
| *XRCC5* | 6 | 0 | 1 | 1 | 2 | 0 | 10 |
| *IRS1* | 2 | 1 | 0 | 0 | 0 | 0 | 3 |
| *INSRA* | 10 | 0 | 0 | 0 | 0 | 0 | 10 |
| *IGF1R* | 8 | 4 | 0 | 0 | 0 | 3 | 15 |
| *FOXO1* | 1 | 0 | 0 | 0 | 0 | 0 | 1 |
| *CEL* | 2 | 0 | 0 | 0 | 0 | 0 | 2 |
| *MGAT5* | Konnte nicht überprüft werden, da in unseren Transkriptkatalogen nicht vorhanden. | | | | | | |
| *C3* | Konnte nicht überprüft werden, da in unseren Transkriptkatalogen nicht vorhanden. | | | | | | |
| **Summe** | 33 | 6 | 1 | 1 | 2 | 3 | 46 |

Anmerkung: Die Zweignummern und Farben korrespondieren mit Abbildung 7. Zweig 5 ist nicht dargestellt, da keine Substitutionen auf ihn entfielen.

Dieses Beispiel demonstriert wie wichtig die Auswahl der Vergleichsspezies in genomweiten Untersuchungen positiver Selektion mit Hinblick auf das zumeist verfolgte Ziel, die genetische Basis bestimmter phänotypischer Eigenschaften zu identifizieren, ist. Wir argumentieren in Manuskript III, dass eine Assoziation zwischen einem PSG und einer solchen Eigenschaft strenggenommen nur dann zulässig ist, wenn ein Schwestertaxon, d.h. eine möglichst nah verwandte Art bzw. Klade in der Analyse repräsentiert ist, die die Eigenschaft nicht teilt. Bspw. war mit Blick auf die Substitutionen an den positiv selektierten Positionen aus Valenzano *et al.* sogar zu erwarten, dass der Großteil von ihnen vor der evolutionären Verkürzung der Lebensspannen auf Zweig 1 stattgefunden hat. Schließlich ist Zweig 1 deutlich länger als die Summe der Längen der Zweige 2 - 6. Vergleichbare Konstellationen lassen sich auch in anderen hochrangig publizierten, genomweiten Untersuchungen positiver Selektion finden, z.B. bei Kim *et al.*, die positive Selektion beim Nacktmull untersucht und als nächstverwandte Vergleichsspezies die evolutionär weit entfernten Mäuse und Ratten verwendet haben (Kim, et al. 2011). Derartiges Vorgehen mag gerade bei älteren Arbeiten dem Umstand geschuldet sein, dass die Sequenzen besser passender Vergleichsspezies nicht öffentlich verfügbar waren und der Aufwand für eine eigene Sequenzierung als zu hoch eingeschätzt wurde. Gerade in Anbetracht der fortschreitenden Verbilligung von Sequenziervorhaben hoffe ich aber mit Manuskript III einen Beitrag dazu geleistet zu haben, dass dem m.E. für die Interpretation der Resultate entscheidenden Aspekt der Auswahl von Vergleichsspezies in Zukunft eine größere Aufmerksamkeit geschenkt wird als bisher.

Auf der einen Seite hat Manuskript III zwar ein konzeptionelles Problem in der Arbeit von Valenzano *et al.* aufgezeigt, auf der anderen Seite hat es aber auch verdeutlicht, dass unser Manuskript II mit der Untersuchung von Zweig 6 nur einen kleinen Teil der evolutionären Lebensspannenverkürzung innerhalb der *Nothobranchius*-Klade abgebildet hat. Auf einigen anzestralen Zweigen der Klade wurde die Lebensspanne mit sehr großer Wahrscheinlichkeit mindestens ebenso stark verringert wie auf dem Zweig des Türkisen Prachtgrundkärpflings. Der Zweig-Positionstest ermöglicht es anzestrale Zweige ebenso zu untersuchen wie terminale, die heute lebende Spezies repräsentieren. Mit Manuskript IV habe ich daher gezielt auf jenen drei anzestralen Zweigen der *Nothobranchius*-Klade nach PSGs gefahndet, auf denen solche evolutionären Anpassungen der Lebensspanne am wahrscheinlichsten sind: Zum habe ich Zweig 2 (Zweig N nach der Notation von Manuskript IV) – also den letzten gemeinsamen Vorfahren aller *Nothobranchius*-Arten – auf positive Selektion untersucht, um den Übergang zur einjährigen Lebensweise zu analysieren (Furness, et al. 2015). Zum anderen habe ich auf den Zweigen 4 und 7 (Zweig FKK bzw. PR nach der Notation von Manuskript IV) nach PSGs gesucht. Alle Nachfahren der durch diese Zweige

repräsentierten anzestralen Spezies leben in einem Gebiet, das Teile Süd- und Zentralmosambiks umfasst. Die Zweige 4 und 7 repräsentieren daher unabhängige Anpassungen an dieselben historischen Klimaveränderungen, die während des durch die Zweige abgedeckten Zeitraums zu verstärkter Trockenheit in diesem Teil Afrikas geführt haben. Daher hielt ich es für möglich, dass die Anpassungen, die zur Verkürzung der Lebensspanne auf den jeweiligen Zweigen beigetragen haben könnten, an ähnlichen biologischer Mechanismen oder im Extremfall sogar auf den selben Genen stattgefunden haben könnten – auf Zweig 4 und 7 im Sinne paralleler und bei beiden im Verhältnis zu Zweig 2 als fortgesetzte Evolution.

Im Ergebnis konnte ich auf allen drei untersuchten Zweigen funktionelle Anreicherungen der PSGs in verschiedenen Bereichen der mitochondrialen Biogenese feststellen, insbesondere für den Komplex I der mitochondrialen Atmungskette (auf Zweig 2 und 7) und für mitochondriale, ribosomale Proteine (MRPs) (Zweig 4 und 7). Es liegen bereits experimentelle Arbeiten vor, die daraufhin hindeuten, dass diese funktionellen Kategorien alternsrelevant sind. So ist die Expressionshöhe von Komplex I Genen negativ assoziiert mit der Lebensspanne von Türkisen Prachtgrundkärpflingen und Mäusen (Miwa, et al. 2014; Baumgart, et al. 2016). Außerdem wurde gezeigt, dass die chemische Inhibition von Komplex I die Lebensspanne von Türkisen Prachtgrundkärpflingen verlängert (Baumgart, et al. 2016) und dass Grönlandwale als langlebigste, bekannte Wirbeltiere eine im Vergleich zu ähnlich großen aber kürzer lebenden Walen niedrigere Komplex I Aktivität haben, die in niedrigeren ROS-Pegeln und geringeren oxidativen Schäden resultiert (Munro, et al. 2013; Gruber, et al. 2015). MRPs wiederum sind entscheidend für die koordinierte Synthese von mitochondrial und kernkodierten Komponenten der Atmungskette (mitonukleäres Gleichgewicht, (Houtkooper, et al. 2013)) – also u.a. auch für den Komplex I. Gleichzeitig gilt diese Koordination, wie in der Einleitung erläutert, selbst als evolutionär konservierter Langlebigkeitsmechanismus. Meine Ergebnisse unterstützen also die genannten experimentellen Arbeiten, die vorschlagen, dass eine kausale Verbindung existiert, zwischen Genen, die das mitonukleäre Gleichgewicht gewährleisten auf der einen und der Länge der Lebensspanne auf der anderen Seite. Die Unterstützung wird dadurch verstärkt, dass die erwähnten Anreicherungen auf mehreren Zweigen gefunden wurden, auf denen die Lebensspanne verkürzt wurde. Darüber hinaus wurden Anzeichen paralleler bzw. fortgesetzter Evolution – d.h. PSGs mit ähnlichen Funktionen – in nahezu allen Bereichen der mitochondrialen Biogenese gefunden: der Transkription mitochondrialer Gene, der Verarbeitung und Stabilisierung mitochondrialer mRNA, der Montage der Komplexe der Atmungskette sowie in den Untereinheiten der Komplexe selbst (siehe Manuskript IV: Abbildung 3). Mittels einer Monte-Carlo-Simulation habe ich zudem gezeigt, dass die über die drei untersuchten Zweige hinweg identifizierten PSGs wesentlich öfter eine Funktion in einem der Bereiche der mitochondrialen Biogenese aufweisen als durch Zufall erklärbar wäre ($p < 10^{-6}$). Auf Einzelgen-Ebene spricht für die These der parallelen Evolution auch der Umstand, dass neun Gene sowohl auf Zweig 4 als auch auf Zweig 7 als positiv selektiert detektiert wurden. Darunter sind bspw. mit *TFB2M* und *POLRMT* zwei von drei Genen, deren Genprodukte den ternären Komplex bilden, der das gesamte mitochondriale Genom transkribiert (Litonin, et al. 2010). *POLRMT* ist zudem eines von mehreren PSGs, die auch bereits auf Zweigen als positiv selektiert identifiziert wurden, die mit Langlebigkeit assoziiert sind – im konkreten Fall von *POLRMT* neben den Zweigen 4 und 7 in Manuskript IV auch auf dem Zweig der Bartfledermaus (Seim, et al. 2013). Das deutet auf die faszinierende Möglichkeit hin, dass durch – vermutlich gegensätzliche – Anpassungen an den gleichen Genen bzw. biologischen Mechanismen sowohl längere als auch kürzere Lebensspannen erreicht werden können. Für diese Möglichkeit spricht zudem, dass auf Ameisen-Zweigen mehrere PSG-Anreicherungen im Bereich der mitochondrialen Biogenese – z.B. für Komplex I und die Untereinheiten des mitochondrialen Ribosoms – identifiziert und in Verbindung mit der 100-fachen Erhöhung der

Lebensspanne in Ameisen in Vergleich zu ihren solitär lebenden Vorfahren gebracht wurde (Roux, et al. 2014).

Die These, dass Anpassungen der Lebensspanne in beide Richtungen z.T. an den gleichen biologischen Mechanismen stattfinden, habe ich – als einen Teilaspekt – in Manuskript V wieder aufgegriffen. Dabei ging ich u.a. von einer Analyse der Genexpression der PSGs während des Alterns im Türkisen Prachtgrundkärpflings aus, die Teil von Manuskript IV war und die ihrerseits auf Ergebnissen aus Manuskript II beruhte, an deren Erzeugung ich nicht beteiligt war. Das wichtigste Resultat dieser Analyse war, dass der Expressionspegel der PSGs auf den drei untersuchten Zweigen während des Alterns signifikant häufiger steigt als fällt (Manuskript IV: Abbildung 5).

## 8. Positive Selektion bei langlebigen Sandgräbern

In Manuskript V habe ich PosiGene dann eingesetzt, um positive Selektion in Zusammenhang mit der Evolution von Langlebigkeit zu untersuchen. Dazu habe ich auf Zweigen von existierenden und anzestralen Nagetierspezies nach PSGs gefahndet, auf den die Lebensspanne wahrscheinlich verlängert wurde. Im Mittelpunkt stand dabei die Familie der Sandgräber (Bathyergidae). Wie im vierten Kapitel geschildert, weisen Grau-, Nackt- und Silbermulle wesentlich höhere Lebensspannen auf als aufgrund des Gewichts dieser Tiere im Verhältnis zu anderen Säugetieren zu erwarten wäre. Außerdem haben sie eine um ein Vielfaches höhere Lebenserwartung als etwa gleich große Nagetiere oder die nahverwandte Rohrratte (Tacutu, et al. 2013). Für diese Analysen habe ich im Unterschied zu den zuvor durchgeführten größtenteils auf öffentliche Daten zurückgriffen. Einige Schlüsselspezies, wie der Nacktmull und zwei Graumullspezies, wurden in unserer Arbeitsgruppe sequenziert. Wie bei Manuskript II haben dabei größtenteils Koautoren die Schritte durchgeführt, die zur Rekonstruktion der CDSs dieser Spezies notwendig waren. Die Assemblierung und CDS-Annotation für den Silbermull und die Rohrratte habe ich selbst übernommen, da für diese genomische Sequenzdaten vorlagen und das in unserer Gruppe entwickelte und zuvor stets genutzte Programm FRAMA (Bens, et al. 2016) für Transkriptomdaten ausgelegt ist.

Insgesamt habe ich mit Manuskript V 341 PSGs auf elf mit Langlebigkeit assoziierten evolutionären Zweigen identifiziert. 20 Gene wurden dabei auf mehreren Zweigen als positiv selektiert detektiert, was wie schon zuvor bei den Prachtgrundkärpflingen darauf hindeutet, dass parallele Evolution möglicherweise nicht nur an ähnlichen biologischen Funktionen, sondern z.T. auch auf den gleichen Genen stattfindet. Bspw. wurde *AMHR2* auf den Zweigen des Nackt- und Blindmulls als PSG detektiert. Die Kinasedomäne des entsprechenden Genprodukts enthält, basierend auf einer Studie mit 33 Säugetieren, die größte Zahl Langlebigkeits-assoziierter Positionen (Semeiks and Grishin 2012). In der gleichen Domäne liegen auch 3 von 8 (Nacktmull) bzw. 2 von 3 (Blindmull) positiv selektierte Positionen. Ich habe zuvor geschrieben, dass in den meisten Fällen Vorwissen aus der Laborforschung nötig ist, um ein PSG als potenziell relevant für die Ausprägung eines bestimmten Phänotyps zu interpretieren. AMHR2 ist m.E. ein Beispiel dafür, dass der experimentellen Forschung durch mehrere unabhängige *in silico* Analysen in einzelnen Fällen sogar gänzlich neue Ansatzpunkte geliefert werden können.

Wie in Kapitel 4 beschrieben, war Manuskript V nicht die erste Untersuchung von Sandgräberzweigen auf positive Selektion. Allerdings haben frühere Arbeiten (Kim, et al. 2011; Fang, Seim, et al. 2014; Davies, et al. 2015) jeweils nur wenige evolutionäre Zweige untersucht und hatten eine geringere phylogenetische Auflösung. So ist Manuskript V bspw. die erste Arbeit, die mit der Rohrratte einen Repräsentanten des Schwestertaxons als Vergleichsspezies einsetzt, was, wie in Manuskript III beschrieben, essentiell für die

Aussagekraft derartiger Analysen ist. Daher sind auch die Ergebnisse der verschiedenen Artikel zur positiven Selektion bei Sandgräbern kaum vergleichbar. Analog zur Konstellation von Valenzano *et al.*/Manuskript II wurden in den verschiedenen Arbeiten *de facto* unterschiedliche phylogenetische Entitäten analysiert.

Meine Ergebnisse von Manuskript V deuten darauf hin, dass die Verlängerung der Lebensspannen auf den analysierten Sandgräberzweigen maßgeblich durch Anpassungen, der antioxidantiellen Verteidigung gegen reaktive Sauerstoffspezies (ROS) und vor allem an vom mTOR-Signalweg regulierten Prozessen wie Translation, Autophagie, Entzündungssystem und mitochondrialer Biogenese erreicht wurden. Für diese in der Einleitung erläuterten, alternsrelevanten Prozesse habe ich in verschiedenen Zusammenhängen Anreicherungen der identifizierten PSGs nachgewiesen sowie konkrete Gene unter positiver Selektion aus der Perspektive der Alternsforschung interpretiert, die m.E. für eventuelle experimentelle Folgestudien in Frage kämen. Es folgen einige, wenige Beispiele.

Als ein Beispiel für die Anpassung der antioxidantiellen Verteidigung möchte ich *TXN* anführen. *TXN* kodiert für ein Oxidoreduktase-Enzym, das u.a. für die Entgiftung von ROS wichtig ist, und wurde auf zwei anzestralen Zweigen als PSG detektiert – auf einem der beiden Zweige als Teil einer Anreicherung für Oxidoreduktaseaktivität. Die Überexpression von TXN verlängert die Lebensspanne von Fliegen (Umeda-Kameyama, et al. 2007) und möglicherweise von Mäusen (Mitsui, et al. 2002; Perez, et al. 2011). Ein Koautor hat die möglichen Auswirkungen der Substitutionen an den positiv selektierten Positionen auf Basis von Homologiemodellen analysiert. Demnach ist es wahrscheinlich, dass durch die Substitutionen die enzymatische Oxidoreduktaseaktivität von TXN und damit auch seine antioxidative Kapazität unter Bedingungen oxidativen Stresses beeinflusst wird (Manuskript V: Abbildung 3). Dies zeigt exemplarisch, wie anhand der von PosiGene erzeugten Alignierungsvisualisierungen biologisch relevante Hypothesen in Bezug auf die positiv selektierten Positionen einzelner PSGs formuliert und mit anderen Methoden weiterverfolgt werden können – in diesem Fall mit einer weiteren *in silico* Methode.

Als Beispiele für Anpassungen am mTOR-Signalweg möchte ich zunächst einen Fall möglicher paralleler Evolution auf Paralogen nennen: *RHEB*, das für den direkten Regulator von mTOR kodiert und auf dem Nacktmullzweig als PSG identifiziert wurde, sowie sein Paralog *RHEBL1*, das auf dem Graumullzweig detektiert wurde. mTOR ist selbst ein zellulärer Schlüsselregulator, dessen Inhibition zu den am besten belegten Lebensspanne-verlängernden Eingriffen zählt (Kenyon 2010; Johnson, et al. 2013). Wie der Einleitung zu entnehmen ist, geht man davon aus, dass Alternsrelevanz von mTOR insbesondere ein Produkt der Regulation jener Prozesse ist, von denen ich in Manuskript V zeige, dass sie auf langlebigkeitsassoziierten Zweigen maßgeblich durch positive Selektion beeinflusst wurden. Mit Blick auf das Immunsystem ist bspw. die Vereinigungsmenge der PSGs über alle Zweige angereichert für die Entzündungs- und Verteidigungsantwort des Organismus. Was zelluläre Selbstreinigung durch Autophagie bzw. das Ubiquitin-Proteasom-System angeht, habe ich u.a. *LAMP2* sowohl auf dem Nackt- als auch auf dem Blindmullzweig als PSG detektiert – ein weiterer Fall möglicher paralleler Evolution auf Genebene. Wird der alternsbedingte Abfall der Expression dieses Rezeptors für Chaperon-vermittelte Autophagie gezielt durch einen genetischen Eingriff ausgeglichen, findet ebenfalls der ansonsten zu verzeichnende Abfall der Autophagie-Aktivität nicht statt (Zhang and Cuervo 2008). Die gleiche Studie hat festgestellt, dass in der Folge die Menge der Zellschäden und die Organfunktion in den untersuchten Mäuselebern auf einem Niveau wiederhergestellt wurde, das vergleichbar mit dem von jungen Mäusen ist. Mit Blick auf mitochondriale Biogenese habe ich u.a. eine Anreicherung von PSGs mit Funktionen in der mitochondrialen Translation auf dem Silbermullzweig festgestellt – darunter u.a. mehrere MRPs. Weitere MRPs, mitochondriale mRNA-Prozessierungsgene und Komplex I Komponenten wurden auf

verschiedenen Zweigen des Baums detektiert (Manuskript V: Tabelle 1). Der Terminationsfaktor der mitochondrialen Transkription *MTERF*, den wir in Manuskript IV auf Zweig 7 als PSG identifiziert hatten, wurde in Manuskript V auf dem Nacktmullzweig detektiert  Die These, dass die mitochondriale Biogenese ein Prozess sein könnte, dessen Anpassung sowohl bei evolutionären Verkürzungen der Lebensspanne (mehrere Prachtgrundkärpflingzweige) als auch bei Verlängerungen der selbigen (Ameisen, Fledermäuse) eine entscheidende Rolle gespielt hat, wurde bereits in Manuskript IV aufgestellt und wird durch die Ergebnisse von Manuskript V gestützt.

Wie bereits festgestellt wurde, war ein weiteres Ergebnis von Manuskript IV, dass der Expressionspegel der PSGs in Prachtgrundkärpflingen im Alter wesentlich häufiger steigt als fällt. Dieser Befund passt zu der in der Einleitung erläuterten evolutionären Alternstheorie der antagonistischen Pleiotropie, die vorhersagt, dass dieselben Gene, die in der Jugend vorteilhaft sind, in späteren Lebensphasen das Altern befördern (Williams 1957). Demzufolge wären die Anpassungen an den PSGs der Prachtgrundkärpflinge vor allem darauf ausgerichtet, ihnen ihr extrem schnelles Wachstum und ihre frühe Fruchtbarkeit zu ermöglichen – der Zusammenhang mit der Biogenese von Mitochondrien, als Dreh- und Angelpunkt des zellulären Energiestoffwechsels und Hauptaspekt von Manuskript IV, liegt auf der Hand. Wenn aber verstärkte Aktivität der PSGs – auf der Ebene der Proteinfunktion oder der Genregulation – zu kürzeren Lebensspannen führt, wäre anzunehmen, dass die Selektion auf Langlebigkeit eher mit einer Dämpfung ihrer Aktivität kompatibel ist, da es i.d.R. leichter ist Schäden zu vermeiden als sie zu reparieren. Diese Hypothese habe ich getestet, indem ich die Vereinigungsmenge der PSGs über die elf zur Langlebigkeit führenden Zweige von Manuskript V hinsichtlich der Richtung der Genexpression während des Alterns beim langlebigen Nacktmull und der kurzlebigen Ratte untersucht habe. Tatsächlich zeigten die PSGs eine signifikante Präferenz für niedrigere Expressionspegel während des Alterns beim Nacktmull und für höhere Expressionspegel bei der Ratte. Mit einer genaueren Betrachtung habe ich zudem nachgewiesen, dass es zu großen Teilen dieselben PSGs sind, die im Alter beim Nacktmull niedriger und bei der Ratte höher exprimiert werden als in der Jugend. Ausgehend von dem Resultat aus Manuskript IV wurde also gezeigt, dass die identifizierten PSGs in lang- und kurzlebigen Arten ebenso klare wie gegensätzliche Expressionsmuster während des Alterns aufweisen, die mit der Alternstheorie der antagonistischen Pleiotropie im Einklang stehen.

Anschließend habe ich untersucht, ob – wie die obigen Resultate nahelegen – Verbindungen hinsichtlich der jeweils betroffenen biologischen Funktionen zwischen positiver Selektion auf der einen und der Veränderung der Genexpression im Alter auf der anderen Seite existieren. Es zeigte sich, dass vier der sechs biologischen Prozesse, die im Nacktmull am stärksten von Änderungen der Genexpression im Alter betroffen sind, sowohl alternsrelevant als auch von positiver Selektion betroffen waren: Zellatmung, Antwort auf oxidativen Stress, Eisenstoffwechsel und Translation. Weiter habe ich nachgewiesen, dass insbesondere jene PSGs, die im langlebigen Nacktmull im Alter höher und in der kurzlebigen Ratte niedriger exprimiert werden, in allen genannten Prozessen überrepräsentiert sind. Dies ist ein weiterer Hinweis darauf, dass diese PSGs auf eine antagonistisch-pleiotrope Weise mit alternsrelevanten Prozessen verbunden sind.

Die Ergebnisse der Manuskripte IV und V unterstützen nicht nur zentrale Vorhersagen der Alternstheorie der antagonistischen Pleiotropie, sondern auch der darauf aufbauenden Hyperfunktionstheorie des Alterns. Diese geht, wie in der Einleitung erläutert, davon aus, dass ein in der Jugend nützliches, vom mTOR-Signalweg gesteuertes Wachstumsprogramm in späteren Lebensphasen Schäden verursacht, die zum Verfall des Organismus führen würden. Übereinstimmend damit deutet meine Arbeit daraufhin, dass die evolutionären Veränderungen der Lebensspannen von Sandgräbern bzw. Prachtgrundkärpflingen

maßgeblich durch Modifikation mTOR-regulierter Schlüsselprozesse wie mRNA-Translation, Autophagie, Entzündungen und mitochondrialer Biogenese verursacht wurden.

## 9. Schlussbemerkungen

Insgesamt unterstreicht meine Arbeit die Relevanz der im vorhergehenden Absatz genannten Theorien und Prozesse für die Alternsforschung. Sie vertritt dabei insbesondere die These, dass die Evolution von Kurz- und Langlebigkeit – mindestens in den untersuchten Spezies, aber möglicherweise auch generell – zu nicht unerheblichen Teilen kausal auf funktionell gegenläufige Anpassungen der gleichen biologischen Prozesse zurückzuführen ist. Was die Methode anbelangt, die meinen Resultaten zu Grunde liegt – die genomweite Suche nach positiver Selektion – so habe ich exemplarisch gezeigt, dass die Auswahl der in die Analyse einbezogenen Spezies entscheidend ist für die Vergleichbarkeit und Aussagekraft entsprechender Studien. Gleichzeitig habe ich diese Methode mit der Veröffentlichung des Programms PosiGene auch weiterentwickelt. PosiGene ist die erste öffentlich verfügbare Software-Lösung, die es ermöglicht PSGs genomweit auf beliebigen, vom Nutzer ausgewählten evolutionären Zweigen zu detektieren. PosiGene verringert den Aufwand und das Wissen erheblich, das zur Durchführung einer solchen Analyse nötig ist. Das Programm liefert verlässliche sowie reproduzierbare Ergebnisse, deren Hauptzweck m.E. darin besteht, dabei zu helfen die genetische Basis spezies- oder kladenspezifischer phänotypischer Eigenschaften aufzuklären. Im Rahmen dieser Arbeit habe ich PosiGene angewendet, um die Suche nach der genetischen Basis des Alterns bzw. besonders kurzer oder langer Lebensspannen zu unterstützen. Klar ist, dass meine Arbeit diese Suche weder beginnt noch beendet. Dennoch liefert sie mit den oben beschriebenen Einsichten zu möglicherweise beteiligten biologischen Prozessen sowie vor allem einer Reihe vielversprechender Genkandidaten wie z.B. *ID3*, *RHEB*, *MTERF*, *POLRMT*, *TXN*, *AMHR2* oder *LAMP2* Ansatzpunkte für Folgestudien. Dabei bieten sich insbesondere die Substitutionen an den von mir identifizierten positiv selektierten Positionen der PSGs für die Untersuchung der Auswirkungen gezielter genetischer Eingriffe mit modernen Methoden wie CRISPR-Cas an. In diesem Sinne sehe ich meine Arbeit als einen Schritt auf dem Weg zu einem besseren Verständnis des Alterns, welches uns hoffentlich in nicht allzu ferner Zukunft erlaubt den alten Menschheitstraum der Verlangsamung dieses Prozesses – und damit eines längeren und gesünderen Lebens – zu verwirklichen.

# Abkürzungsverzeichnis

| | |
|---|---|
| bspw. | beispielsweise |
| ca. | circa |
| CDS | *Protein coding sequence* (proteinkodierende Sequenz) |
| d.h. | das heißt |
| DNA | *Deoxyribonucleic acid* (Desoxyribonukleinsäure) |
| *et al.* | *et alii* (und andere) |
| etc. | *et cetera* (und die übrigen Dinge) |
| ggf. | gegebenenfalls |
| HIV | Humanes Immundefizienz-Virus |
| HZ | Hintergrundzweig |
| IF | *Impact factor* (Einflussfaktor) |
| InDels | Insertionen/Deletionen |
| Kb | Kilobasen |
| mRNA | *messenger RNA* (Boten-RNA) |
| Mb | Megabasen |
| PSG | positiv selektiertes Gen |
| RNA | *Ribonucleic acid* (Ribonukleinsäure) |
| u.a. | unter anderem |
| usw. | und so weiter |
| u.v.m. | und vieles mehr |
| VZ | Vordergrundzweig |
| z.B. | zum Beispiel |

# Literaturverzeichnis

Alfoldi J, Lindblad-Toh K. 2013. Comparative genomics as a tool to understand evolution and disease. Genome Res 23:1063-1068.
Bakewell MA, Shi P, Zhang J. 2007. More genes underwent positive selection in chimpanzee evolution than in human evolution. Proc Natl Acad Sci U S A 104:7489-7494.
Bartke A. 2012. Healthy aging: is smaller better? - a mini-review. Gerontology 58:337-343.
Baumgart M, Groth M, Priebe S, Savino A, Testa G, Dix A, Ripa R, Spallotta F, Gaetano C, Ori M, et al. 2014. RNA-seq of the aging brain in the short-lived fish N. furzeri - conserved pathways and novel genes associated with neurogenesis. Aging Cell 13:965-974.
Baumgart M, Priebe S, Groth M, Hartmann N, Menzel U, Pandolfini L, Koch P, Felder M, Ristow M, Englert C, et al. 2016. Longitudinal RNA-Seq Analysis of Vertebrate Aging Identifies Mitochondrial Complex I as a Small-Molecule-Sensitive Modifier of Lifespan. Cell Syst 2:122-132.
Bennett NC, Faulkes CG. 2000. African Mole-Rats: Ecology and Eusociality. Cambridge University Press.
Bens M, Sahm A, Groth M, Jahn N, Morhart M, Holtze S, Hildebrandt TB, Platzer M, Szafranski K. 2016. FRAMA: from RNA-seq data to annotated mRNA assemblies. BMC Genomics 17:54.
Biswas S, Akey JM. 2006. Genomic insights into positive selection. Trends Genet 22:437-446.
Blagosklonny MV. 2012. Answering the ultimate question "what is the proximal cause of aging?". Aging (Albany NY) 4:861-877.
Blanga-Kanfi S, Miranda H, Penn O, Pupko T, DeBry RW, Huchon D. 2009. Rodent phylogeny revised: analysis of six nuclear genes from all major rodent clades. BMC Evol Biol 9:71.
Bratic A, Larsson NG. 2013. The role of mitochondria in aging. J Clin Invest 123:951-957.
Braunseis F, Deutsch T, Frese T, Sandholzer H. 2012. The risk for nursing home admission (NHA) did not change in ten years--a prospective cohort study with five-year follow-up. Arch Gerontol Geriatr 54:e63-67.
Broer L, Buchman AS, Deelen J, Evans DS, Faul JD, Lunetta KL, Sebastiani P, Smith JA, Smith AV, Tanaka T, et al. 2015. GWAS of longevity in CHARGE consortium confirms APOE and FOXO3 candidacy. J Gerontol A Biol Sci Med Sci 70:110-118.
Buffenstein R. 2008. Negligible senescence in the longest living rodent, the naked mole-rat: insights from a successfully aging species. J Comp Physiol B 178:439-445.
Burda H. 2001. Determinants of the distribution and radiation of African mole-rats (Bathyergidae, Rodentia): Ecology or geography? African Small Mammals:261-277.
Burda H, Honeycutt RL, Begall S, Locker-Grütjen O, Scharff A. 2000. Are naked and common mole-rats eusocial and if so, why? Behavioral Ecology and Sociobiology 47:293-303.
Butler RN, Miller RA, Perry D, Carnes BA, Williams TF, Cassel C, Brody J, Bernard MA, Partridge L, Kirkwood T, et al. 2008. New model of health promotion and disease prevention for the 21st century. BMJ 337:a399.
Carlini DB, Stephan W. 2003. In vivo introduction of unpreferred synonymous codons into the Drosophila Adh gene results in reduced levels of ADH protein. Genetics 163:239-243.
Carmona JJ, Michan S. 2016. Biology of Healthy Aging and Longevity. Rev Invest Clin 68:7-16.
Cellerino A, Valenzano DR, Reichard M. 2016. From the bush to the bench: the annual Nothobranchius fishes as a new model system in biology. Biol Rev Camb Philos Soc 91:511-533.
Chandrasekaran A, Idelchik MD, Melendez JA. 2017. Redox control of senescence and age-related disease. Redox Biol 11:91-102.
Chen J, Astle CM, Harrison DE. 2000. Genetic regulation of primitive hematopoietic stem cell senescence. Exp Hematol 28:442-450.
Chen JQ, Wu Y, Yang H, Bergelson J, Kreitman M, Tian D. 2009. Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. Mol Biol Evol 26:1523-1531.
Chen X, Su Z, Dam P, Palenik B, Xu Y, Jiang T. 2004. Operon prediction by comparative genomics: an application to the Synechococcus sp. WH8102 genome. Nucleic Acids Res 32:2147-2157.
Christensen K, Doblhammer G, Rau R, Vaupel JW. 2009. Ageing populations: the challenges ahead. Lancet 374:1196-1208.
Chung HY, Cesari M, Anton S, Marzetti E, Giovannini S, Seo AY, Carter C, Yu BP, Leeuwenburgh C. 2009. Molecular inflammation: underpinnings of aging and age-related diseases. Ageing Res Rev 8:18-30.
Collins TM, Wimberger PH, Naylor J. 1994. Compositional Bias, Character-State Bias, and Character-State Reconstruction Using Parsimony Systematic Biology 43:482-496.
Copeland JM, Cho J, Lo T, Jr., Hur JH, Bahadorani S, Arabyan T, Rabie J, Soh J, Walker DW. 2009. Extension of Drosophila life span by RNAi of the mitochondrial respiratory chain. Curr Biol 19:1591-1598.
Correia-Melo C, Passos JF. 2015. Mitochondria: Are they causal players in cellular senescence? Biochim Biophys Acta 1847:1373-1379.
Costa JT, Fitzgerald TD. 1996. Developments in social terminology: semantic battles in a conceptual war. Trends Ecol Evol 11:285-289.

Crimmins EM. 2015. Lifespan and Healthspan: Past, Present, and Promise. Gerontologist 55:901-911.
Crimmins EM, Beltran-Sanchez H. 2011. Mortality and morbidity trends: is there compression of morbidity? J Gerontol B Psychol Sci Soc Sci 66:75-86.
Cunningham GM, Roman MG, Flores LC, Hubbard GB, Salmon AB, Zhang Y, Gelfond J, Ikeno Y. 2015. The paradoxical role of thioredoxin on oxidative stress and aging. Arch Biochem Biophys 576:32-38.
Dammann P, Sumbera R, Massmann C, Scherag A, Burda H. 2011. Extended longevity of reproductives appears to be common in Fukomys mole-rats (Rodentia, Bathyergidae). PLoS One 6:e18757.
Darwin C. 1859. On the origin of species by means of natural selection.
Davalli P, Mitic T, Caporali A, Lauriola A, D'Arca D. 2016. ROS, Cell Senescence, and Novel Molecular Mechanisms in Aging and Age-Related Diseases. Oxid Med Cell Longev 2016:3565127.
Davies KT, Bennett NC, Tsagkogeorga G, Rossiter SJ, Faulkes CG. 2015. Family Wide Molecular Adaptations to Underground Life in African Mole-Rats Revealed by Phylogenomic Analysis. Mol Biol Evol 32:3089-3107.
de Magalhaes JP. 2015. The big, the bad and the ugly: Extreme animals as inspiration for biomedical research. EMBO Rep 16:771-776.
de Magalhaes JP, Costa J, Church GM. 2007. An analysis of the relationship between metabolism, developmental schedules, and longevity using phylogenetic independent contrasts. J Gerontol A Biol Sci Med Sci 62:149-160.
Defrance M, Touzet H. 2006. Predicting transcription factor binding sites using local over-representation and comparative genomics. BMC Bioinformatics 7:396.
Di Cicco E, Tozzini ET, Rossi G, Cellerino A. 2011. The short-lived annual fish Nothobranchius furzeri shows a typical teleost aging process reinforced by high incidence of age-dependent neoplasias. Exp Gerontol 46:249-256.
Dillin A, Hsu AL, Arantes-Oliveira N, Lehrer-Graiwer J, Hsin H, Fraser AG, Kamath RS, Ahringer J, Kenyon C. 2002. Rates of behavior and aging specified by mitochondrial function during development. Science 298:2398-2401.
Dong X, Milholland B, Vijg J. 2016. Evidence for a limit to human lifespan. Nature 538:257-259.
Dorn A, Musilova Z, Platzer M, Reichwald K, Cellerino A. 2014. The strange case of East African annual fishes: aridification correlates with diversification for a savannah aquatic group? BMC Evol Biol 14:210.
Edrey YH, Salmon AB. 2014. Revisiting an age-old question regarding oxidative stress. Free Radic Biol Med 71:368-378.
Fang X, Nevo E, Han L, Levanon EY, Zhao J, Avivi A, Larkin D, Jiang X, Feranchuk S, Zhu Y, et al. 2014. Genome-wide adaptive complexes to underground stresses in blind mole rats Spalax. Nat Commun 5:3966.
Fang X, Seim I, Huang Z, Gerashchenko MV, Xiong Z, Turanov AA, Zhu Y, Lobanov AV, Fan D, Yim SH, et al. 2014. Adaptations to a subterranean environment and longevity revealed by the analysis of mole rat genomes. Cell Rep 8:1354-1364.
Farrelly C. 2008. Has the time come to take on time itself? BMJ 337:a414.
Felsenstein J. 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J Mol Evol 17:368-376.
Filomeni G, De Zio D, Cecconi F. 2015. Oxidative stress and autophagy: the clash between damage and metabolic needs. Cell Death Differ 22:377-388.
Finch CE. 2009. Update on slow aging and negligible senescence--a mini-review. Gerontology 55:307-313.
Fletcher W, Yang Z. 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol Biol Evol 27:2257-2267.
Fontana L, Partridge L, Longo VD. 2010. Extending healthy life span--from yeast to humans. Science 328:321-326.
Fries JF, Bruce B, Chakravarty E. 2011. Compression of morbidity 1980-2011: a focused review of paradigms and progress. J Aging Res 2011:261702.
Fu W, Akey JM. 2013. Selection and adaptation in the human genome. Annu Rev Genomics Hum Genet 14:467-489.
Furness AI, Reznick DN, Springer MS, Meredith RW. 2015. Convergent evolution of alternative developmental trajectories associated with diapause in African and South American killifish. Proc Biol Sci 282.
Fushan AA, Turanov AA, Lee SG, Kim EB, Lobanov AV, Yim SH, Buffenstein R, Lee SR, Chang KT, Rhee H, et al. 2015. Gene expression defines natural changes in mammalian lifespan. Aging Cell 14:352-365.
Gaffney DJ, Keightley PD. 2005. The scale of mutational variation in the murid genome. Genome Res 15:1086-1094.
Galenkamp H, Braam AW, Huisman M, Deeg DJ. 2013. Seventeen-year time trend in poor self-rated health in older adults: changing contributions of chronic diseases and disability. Eur J Public Health 23:511-517.
Gaya-Vidal M, Alba MM. 2014. Uncovering adaptive evolution in the human lineage. BMC Genomics 15:599.
Gems D, Partridge L. 2013. Genetics of longevity in model organisms: debates and paradigm shifts. Annu Rev Physiol 75:621-644.
Gharib WH, Robinson-Rechavi M. 2013. The branch-site test of positive selection is surprisingly robust but lacks power under synonymous substitution saturation and variation in GC. Mol Biol Evol 30:1675-1686.

Goldman DP, Cutler D, Rowe JW, Michaud PC, Sullivan J, Peneva D, Olshansky SJ. 2013. Substantial health and economic returns from delayed aging may warrant a new focus for medical research. Health Aff (Millwood) 32:1698-1705.
Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. Mol Biol Evol 11:725-736.
Gorbunova V, Seluanov A, Zhang Z, Gladyshev VN, Vijg J. 2014. Comparative genetics of longevity and cancer: insights from long-lived rodents. Nat Rev Genet 15:531-540.
Gruber H, Wessels W, Boynton P, Xu J, Wohlgemuth S, Leeuwenburgh C, Qi W, Austad SN, Schaible R, Philipp EE. 2015. Age-related cellular changes in the long-lived bivalve A. islandica. Age (Dordr) 37:90.
Hands SL, Proud CG, Wyttenbach A. 2009. mTOR's role in ageing: protein synthesis or autophagy? Aging (Albany NY) 1:586-597.
Harman D. 2001. Aging: overview. Ann N Y Acad Sci 928:1-21.
Healy K. 2015. Eusociality but not fossoriality drives longevity in small mammals. Proc Biol Sci 282:20142917.
Healy K, Guillerme T, Finlay S, Kane A, Kelly SB, McClean D, Kelly DJ, Donohue I, Jackson AL, Cooper N. 2014. Ecology and mode-of-life explain lifespan variation in birds and mammals. Proc Biol Sci 281:20140298.
Hekimi S, Wang Y, Noe A. 2016. Mitochondrial ROS and the Effectors of the Intrinsic Apoptotic Pathway in Aging Cells: The Discerning Killers! Front Genet 7:161.
Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. Nat Rev Genet 12:756-766.
Hofer-Warbinek R, Schmid JA, Mayer H, Winsauer G, Orel L, Mueller B, Wiesner C, Binder BR, de Martin R. 2004. A highly conserved proapoptotic gene, IKIP, located next to the APAF1 gene locus, is regulated by p53. Cell Death Differ 11:1317-1325.
Hofmann JW, Zhao X, De Cecco M, Peterson AL, Pagliaroli L, Manivannan J, Hubbard GB, Ikeno Y, Zhang Y, Feng B, et al. 2015. Reduced expression of MYC increases longevity and enhances healthspan. Cell 160:477-488.
Hongo JA, de Castro GM, Cintra LC, Zerlotini A, Lobo FP. 2015. POTION: an end-to-end pipeline for positive Darwinian selection detection in genome-scale data through phylogenetic comparison of protein-coding genes. BMC Genomics 16:567.
Hou L, Guo L, Wang C, Wang C. 2016. Genome sequence of Candida versatilis and comparative analysis with other yeast. J Ind Microbiol Biotechnol 43:1131-1138.
Houtkooper RH, Mouchiroud L, Ryu D, Moullan N, Katsyuba E, Knott G, Williams RW, Auwerx J. 2013. Mitonuclear protein imbalance as a conserved longevity mechanism. Nature 497:451-457.
Hudson RR, Kreitman M, Aguade M. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics 116:153-159.
Hughes KA, Reynolds RM. 2005. Evolutionary and mechanistic theories of aging. Annu Rev Entomol 50:421-445.
Jarvis JU, Bennett NC. 1993. Eusociality has evolved independently in two genera of bathyergid mole-rats — but occurs in no other subterranean mammal. Behavioral Ecology and Sociobiology 33:253-260.
Ji XH, Vlasak J, Zhou LW, Wu F, Dai YC. 2017. Phylogeny and diversity of Fomitiporella (Hymenochaetales, Basidiomycota). Mycologia:1-15.
Jobson RW, Nabholz B, Galtier N. 2010. An evolutionary genome scan for longevity-related natural selection in mammals. Mol Biol Evol 27:840-847.
Johnson SC, Rabinovitch PS, Kaeberlein M. 2013. mTOR is a key modulator of ageing and age-related disease. Nature 493:338-345.
Jones OR, Scheuerlein A, Salguero-Gomez R, Camarda CG, Schaible R, Casper BB, Dahlgren JP, Ehrlen J, Garcia MB, Menges ES, et al. 2014. Diversity of ageing across the tree of life. Nature 505:169-173.
Jubb RA. 1981. Nothobranchius.
Kaeberlein M. 2013. Longevity and aging. F1000Prime Rep 5:5.
Keane M, Semeiks J, Webb AE, Li YI, Quesada V, Craig T, Madsen LB, van Dam S, Brawand D, Marques PI, et al. 2015. Insights into the evolution of longevity from the bowhead whale genome. Cell Rep 10:112-122.
Keller L, Genoud M. 1997. Extraordinary lifespans in ants: a test of evolutionary theories of aging. Nature 389:958-960.
Kenyon CJ. 2010. The genetics of ageing. Nature 464:504-512.
Kim EB, Fang X, Fushan AA, Huang Z, Lobanov AV, Han L, Marino SM, Sun X, Turanov AA, Yang P, et al. 2011. Genome sequencing reveals insights into physiology and longevity of the naked mole rat. Nature 479:223-227.
Kim GH, Kim JE, Rhie SJ, Yoon S. 2015. The Role of Oxidative Stress in Neurodegenerative Diseases. Exp Neurobiol 24:325-340.
Kinsella RJ, Kahari A, Haider S, Zamora J, Proctor G, Spudich G, Almeida-King J, Staines D, Derwent P, Kerhornou A, et al. 2011. Ensembl BioMarts: a hub for data retrieval across taxonomic space. Database (Oxford) 2011:bar030.
Kosakovsky Pond SL, Frost SD. 2005. Not so different after all: a comparison of methods for detecting amino acid sites under selection. Mol Biol Evol 22:1208-1222.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six Mammalian genomes. PLoS Genet 4:e1000144.
Kotiadis VN, Duchen MR, Osellame LD. 2014. Mitochondrial quality control and communications with the nucleus are important in maintaining mitochondrial function and cell health. Biochim Biophys Acta 1840:1254-1265.
Krieglstein K, Miyazono K, ten Dijke P, Unsicker K. 2012. TGF-beta in aging and disease. Cell Tissue Res 347:5-9.
Labunskyy VM, Gladyshev VN. 2013. Role of reactive oxygen species-mediated signaling in aging. Antioxid Redox Signal 19:1362-1372.
Lagunas-Rangel FA, Chavez-Valencia V. 2017. Learning of nature: The curious case of the naked mole rat. Mech Ageing Dev 164:76-81.
Lamming DW, Ye L, Katajisto P, Goncalves MD, Saitoh M, Stevens DM, Davis JG, Salmon AB, Richardson A, Ahima RS, et al. 2012. Rapamycin-induced insulin resistance is mediated by mTORC2 loss and uncoupled from longevity. Science 335:1638-1643.
Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, et al. 2001. Initial sequencing and analysis of the human genome. Nature 409:860-921.
Laplante M, Sabatini DM. 2012. mTOR signaling in growth control and disease. Cell 149:274-293.
Lewontin RC, Krakauer J. 1973. Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74:175-195.
Li L, Stoeckert CJ, Jr., Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13:2178-2189.
Licastro F, Candore G, Lio D, Porcellini E, Colonna-Romano G, Franceschi C, Caruso C. 2005. Innate immunity and inflammation in ageing: a key for understanding age-related diseases. Immun Ageing 2:8.
Lionaki E, Gkikas I, Tavernarakis N. 2016. Differential Protein Distribution between the Nucleus and Mitochondria: Implications in Aging. Front Genet 7:162.
Litonin D, Sologub M, Shi Y, Savkina M, Anikin M, Falkenberg M, Gustafsson CM, Temiakov D. 2010. Human mitochondrial transcription revisited: only TFAM and TFB2M are required for transcription of the mitochondrial genes in vitro. J Biol Chem 285:18129-18133.
Longo VD, Antebi A, Bartke A, Barzilai N, Brown-Borg HM, Caruso C, Curiel TJ, de Cabo R, Franceschi C, Gems D, et al. 2015. Interventions to Slow Aging in Humans: Are We Ready? Aging Cell 14:497-510.
Lopez-Otin C, Blasco MA, Partridge L, Serrano M, Kroemer G. 2013. The hallmarks of aging. Cell 153:1194-1217.
Loughran NB, Hinde S, McCormick-Hill S, Leidal KG, Bloomberg S, Loughran ST, O'Connor B, O'Fagain C, Nauseef WM, O'Connell MJ. 2012. Functional consequence of positive selection revealed through rational mutagenesis of human myeloperoxidase. Mol Biol Evol 29:2039-2046.
Loytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science 320:1632-1635.
Lucas-Sanchez A, Almaida-Pagan PF, Mendiola P, de Costa J. 2014. Nothobranchius as a model for aging studies. A review. Aging Dis 5:281-291.
Mallick S, Gnerre S, Muller P, Reich D. 2009. The difficulty of avoiding false positives in genome scans for natural selection. Genome Res 19:922-933.
Margolis-Nunno H, Halpern-Sebold L, Schreibman MP. 1986. Immunocytochemical changes in serotonin in the forebrain and pituitary of aging fish. Neurobiol Aging 7:17-21.
Markova-Raina P, Petrov D. 2011. High sensitivity to aligner and high rate of false positives in the estimates of positive selection in the 12 Drosophila genomes. Genome Res 21:863-874.
Martinez-Lopez N, Athonvarangkul D, Singh R. 2015. Autophagy and aging. Adv Exp Med Biol 847:73-87.
McClellan DA. 2013. Directional Darwinian Selection in proteins. BMC Bioinformatics 14 Suppl 13:S6.
McCormick MA, Tsai SY, Kennedy BK. 2011. TOR and ageing: a complex pathway for a complex process. Philos Trans R Soc Lond B Biol Sci 366:17-27.
McDonald JH, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. Nature 351:652-654.
Medvedev ZA. 1990. An attempt at a rational classification of theories of ageing. Biol Rev Camb Philos Soc 65:375-398.
Meng J, Lv Z, Qiao X, Li X, Li Y, Zhang Y, Chen C. 2017. The decay of Redox-stress Response Capacity is a substantive characteristic of aging: Revising the redox theory of aging. Redox Biol 11:365-374.
Miller RA, Harrison DE, Astle CM, Fernandez E, Flurkey K, Han M, Javors MA, Li X, Nadon NL, Nelson JF, et al. 2014. Rapamycin-mediated lifespan increase in mice is dose and sex dependent and metabolically distinct from dietary restriction. Aging Cell 13:468-477.
Mitsui A, Hamuro J, Nakamura H, Kondo N, Hirabayashi Y, Ishizaki-Koizumi S, Hirakawa T, Inoue T, Yodoi J. 2002. Overexpression of human thioredoxin in transgenic mice controls oxidative stress and life span. Antioxid Redox Signal 4:693-696.
Mittal M, Siddiqui MR, Tran K, Reddy SP, Malik AB. 2014. Reactive oxygen species in inflammation and tissue injury. Antioxid Redox Signal 20:1126-1167.

Miwa S, Jow H, Baty K, Johnson A, Czapiewski R, Saretzki G, Treumann A, von Zglinicki T. 2014. Low abundance of the matrix arm of complex I in mitochondria predicts longevity in mice. Nat Commun 5:3837.
Miyata T, Yasunaga T. 1980. Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application. J Mol Evol 16:23-36.
Munro D, Pichaud N, Paquin F, Kemeid V, Blier PU. 2013. Low hydrogen peroxide production in mitochondria of the long-lived Arctica islandica: underlying mechanisms for slow aging. Aging Cell 12:584-592.
Near TJ, Eytan RI, Dornburg A, Kuhn KL, Moore JA, Davis MP, Wainwright PC, Friedman M, Smith WL. 2012. Resolution of ray-finned fish phylogeny and timing of diversification. Proc Natl Acad Sci U S A 109:13698-13703.
Nei M, Gojobori T. 1986. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3:418-426.
Ng'oma E, Reichwald K, Dorn A, Wittig M, Balschun T, Franke A, Platzer M, Cellerino A. 2014. The age related markers lipofuscin and apoptosis show different genetic architecture by QTL mapping in short-lived Nothobranchius fish. Aging (Albany NY) 6:468-480.
Nielsen R. 2001. Statistical tests of selective neutrality in the age of genomics. Heredity (Edinb) 86:641-647.
Nozawa M, Suzuki Y, Nei M. 2009. Reliabilities of identifying positive selection by the branch-site and the site-prediction methods. Proc Natl Acad Sci U S A 106:6700-6705.
Oeppen J, Vaupel JW. 2002. Demography. Broken limits to life expectancy. Science 296:1029-1031.
Olshansky SJ. 2015. Has the Rate of Human Aging Already Been Modified? Cold Spring Harb Perspect Med 5.
Onken B, Driscoll M. 2010. Metformin induces a dietary restriction-like state and the oxidative stress response to extend C. elegans Healthspan via AMPK, LKB1, and SKN-1. PLoS One 5:e8758.
Ori A, Toyama BH, Harris MS, Bock T, Iskar M, Bork P, Ingolia NT, Hetzer MW, Beck M. 2015. Integrated Transcriptome and Proteome Analyses Reveal Organ-Specific Proteome Deterioration in Old Rats. Cell Syst 1:224-237.
Palacios T, Solari C, Bains W. 2015. Prosper and Live Long: Productive Life Span Tracks Increasing Overall Life Span Over Historical Time among Privileged Worker Groups. Rejuvenation Res 18:234-244.
Parker MG, Thorslund M. 2007. Health trends in the elderly population: getting better and getting worse. Gerontologist 47:150-158.
Parmley JL, Chamary JV, Hurst LD. 2006. Evidence for purifying selection against synonymous mutations in mammalian exonic splicing enhancers. Mol Biol Evol 23:301-309.
Perez VI, Cortez LA, Lew CM, Rodriguez M, Webb CR, Van Remmen H, Chaudhuri A, Qi W, Lee S, Bokov A, et al. 2011. Thioredoxin 1 overexpression extends mainly the earlier part of life span in mice. J Gerontol A Biol Sci Med Sci 66:1286-1299.
Petersen L, Bollback JP, Dimmic M, Hubisz M, Nielsen R. 2007. Genes under positive selection in Escherichia coli. Genome Res 17:1336-1343.
Petersen M, Meusemann K, Donath A, Dowling D, Liu S, Peters RS, Podsiadlowski L, Vasilikopoulos A, Zhou X, Misof B, et al. 2017. Orthograph: a versatile tool for mapping coding nucleotide sequences to clusters of orthologous genes. BMC Bioinformatics 18:111.
Phelan JP, Austad SN. 1989. Natural selection, dietary restriction, and extended longevity. Growth Dev Aging 53:4-6.
Pitt JN, Kaeberlein M. 2015. Why is aging conserved and what can we do about it? PLoS Biol 13:e1002131.
Platzer M, Reichwald K, Hartmann N, Cellerino A, Englert C. 2011. Der annuelle Türkise Prachtgrundkärpfling Nothobranchius furzeri als neues Modell für die Alternsforschung. Altersmedizin aktuell 21.
Pond SL, Frost SD, Muse SV. 2005. HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676-679.
Privman E, Penn O, Pupko T. 2012. Improving the performance of positive selection inference by filtering unreliable alignment regions. Mol Biol Evol 29:1-5.
Reichwald K, Petzold A, Koch P, Downie BR, Hartmann N, Pietsch S, Baumgart M, Chalopin D, Felder M, Bens M, et al. 2015. Insights into Sex Chromosome Evolution and Aging from the Genome of a Short-Lived Fish. Cell 163:1527-1538.
Ristow M, Schmeisser S. 2011. Extending life span by increasing oxidative stress. Free Radic Biol Med 51:327-336.
Roux J, Privman E, Moretti S, Daub JT, Robinson-Rechavi M, Keller L. 2014. Patterns of positive selection in seven ant genomes. Mol Biol Evol 31:1661-1685.
Rowe DL, Dunn KA, Adkins RM, Honeycutt RL. 2010. Molecular clocks keep dispersal hypotheses afloat: evidence for trans-Atlantic rafting by rodents. Journal of Biogeography 37:305-324.
Rubinsztein DC, Marino G, Kroemer G. 2011. Autophagy and aging. Cell 146:682-695.
Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, Gabriel SB, Platko JV, Patterson NJ, McDonald GJ, et al. 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419:832-837.
Saez I, Vilchez D. 2014. The Mechanistic Links Between Proteasome Activity, Aging and Age-related Diseases. Curr Genomics 15:38-51.

Sakharkar MK, Chow VT, Kangueane P. 2004. Distributions of exons and introns in the human genome. In Silico Biol 4:387-393.
Salminen A, Kaarniranta K, Kauppinen A. 2012. Inflammaging: disturbed interplay between autophagy and inflammasomes. Aging (Albany NY) 4:166-175.
Sawyer SL, Wu LI, Emerman M, Malik HS. 2005. Positive selection of primate TRIM5alpha identifies a critical species-specific retroviral restriction domain. Proc Natl Acad Sci U S A 102:2832-2837.
Schieber M, Chandel NS. 2014. ROS function in redox signaling and oxidative stress. Curr Biol 24:R453-462.
Schneider A, Souvorov A, Sabath N, Landan G, Gonnet GH, Graur D. 2009. Estimates of positive Darwinian selection are inflated by errors in sequencing, annotation, and alignment. Genome Biol Evol 1:114-118.
Schoeni RF, Freedman VA, Martin LG. 2008. Why is late-life disability declining? Milbank Q 86:47-89.
Seim I, Fang X, Xiong Z, Lobanov AV, Huang Z, Ma S, Feng Y, Turanov AA, Zhu Y, Lenz TL, et al. 2013. Genome analysis reveals insights into physiology and longevity of the Brandt's bat Myotis brandtii. Nat Commun 4:2212.
Semeiks J, Grishin NV. 2012. A method to find longevity-selected positions in the mammalian proteome. PLoS One 7:e38595.
Seney ML, Kelly DA, Goldman BD, Sumbera R, Forger NG. 2009. Social structure predicts genital morphology in African mole-rats. PLoS One 4:e7477.
Sepulchre P, Ramstein G, Fluteau F, Schuster M, Tiercelin JJ, Brunet M. 2006. Tectonic uplift and Eastern Africa aridification. Science 313:1419-1423.
Shen J, Tower J. 2009. Programmed cell death and apoptosis in aging and life span regulation. Discov Med 8:223-226.
Shen YY, Liang L, Zhu ZH, Zhou WP, Irwin DM, Zhang YP. 2010. Adaptive evolution of energy metabolism genes and the origin of flight in bats. Proc Natl Acad Sci U S A 107:8666-8671.
Sheng Z, Schramm CA, Connors M, Morris L, Mascola JR, Kwong PD, Shapiro L. 2016. Effects of Darwinian Selection and Mutability on Rate of Broadly Neutralizing Antibody Evolution during HIV-1 Infection. PLoS Comput Biol 12:e1004940.
Sierra F, Hadley E, Suzman R, Hodes R. 2009. Prospects for life span extension. Annu Rev Med 60:457-469.
Statistisches Bundesamt. 2016. Sterbetafel 2013/2015 - Methoden- und Ergebnisbericht zur laufenden Berechnung von Periodensterbetafeln für Deutschland und die Bundesländer
Steegenga WT, Boekschoten MV, Lute C, Hooiveld GJ, de Groot PJ, Morris TJ, Teschendorff AE, Butcher LM, Beck S, Muller M. 2014. Genome-wide age-related changes in DNA methylation and gene expression in human PBMCs. Age (Dordr) 36:9648.
Stoletzki N. 2008. Conflicting selection pressures on synonymous codon use in yeast suggest selection on mRNA secondary structures. BMC Evol Biol 8:224.
Strehler B. 1977. Times, Cells and Aging.12-15.
Tacutu R, Craig T, Budovsky A, Wuttke D, Lehmann G, Taranukha D, Costa J, Fraifeld VE, de Magalhaes JP. 2013. Human Ageing Genomic Resources: integrated databases and tools for the biology and genetics of ageing. Nucleic Acids Res 41:D1027-1033.
Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585-595.
Talavera G, Castresana J. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564-577.
Tilstra JS, Robinson AR, Wang J, Gregg SQ, Clauson CL, Reay DP, Nasto LA, St Croix CM, Usas A, Vo N, et al. 2012. NF-kappaB inhibition delays DNA damage-induced senescence and aging in mice. J Clin Invest 122:2601-2612.
Tozzini ET, Dorn A, Ng'oma E, Polacik M, Blazek R, Reichwald K, Petzold A, Watters B, Reichard M, Cellerino A. 2013. Parallel evolution of senescence in annual fishes in response to extrinsic mortality. BMC Evol Biol 13:77.
Trauth MH, Maslin MA, Deino A, Strecker MR. 2005. Late Cenozoic moisture history of East Africa. Science 309:2051-2053.
Umeda-Kameyama Y, Tsuda M, Ohkura C, Matsuo T, Namba Y, Ohuchi Y, Aigaki T. 2007. Thioredoxin suppresses Parkin-associated endothelin receptor-like receptor-induced neurotoxicity and extends longevity in Drosophila. J Biol Chem 282:11180-11187.
Valdesalici S, Cellerino A. 2003. Extremely short lifespan in the annual fish Nothobranchius furzeri. Proc Biol Sci 270 Suppl 2:S189-191.
Valenzano DR, Aboobaker A, Seluanov A, Gorbunova V. 2017. Non-canonical aging model systems and why we need them. EMBO J.
Valenzano DR, Benayoun BA, Singh PP, Zhang E, Etter PD, Hu CK, Clement-Ziza M, Willemsen D, Cui R, Harel I, et al. 2015. The African Turquoise Killifish Genome Provides Insights into Evolution and Genetic Architecture of Lifespan. Cell 163:1539-1554.
Vaupel JW. 2010. Biodemography of human ageing. Nature 464:536-542.

Vaupel JW, Lundstrom H. 1994. Longer life expectancy? Evidence from sweden of reductions in mortality rates at advanced ages. WiseDA, ed. Studies in the Economics of Aging:79-102.
Villanueva-Canas JL, Laurie S, Alba MM. 2013. Improving genome-wide scans of positive selection by using protein isoforms of similar length. Genome Biol Evol 5:457-467.
Vinogradov AE. 2003. DNA helix: the importance of being GC-rich. Nucleic Acids Res 31:1838-1844.
Vistoli G, De Maddis D, Cipak A, Zarkovic N, Carini M, Aldini G. 2013. Advanced glycoxidation and lipoxidation end products (AGEs and ALEs): an overview of their mechanisms of formation. Free Radic Res 47 Suppl 1:3-27.
Wang ET, Kodama G, Baldi P, Moyzis RK. 2006. Global landscape of recent inferred Darwinian selection for Homo sapiens. Proc Natl Acad Sci U S A 103:135-140.
Wang H, Yang H, Shivalila CS, Dawlaty MM, Cheng AW, Zhang F, Jaenisch R. 2013. One-step generation of mice carrying mutations in multiple genes by CRISPR/Cas-mediated genome engineering. Cell 153:910-918.
Webb AE, Gerek ZN, Morgan CC, Walsh TA, Loscher CE, Edwards SV, O'Connell MJ. 2015. Adaptive Evolution as a Predictor of Species-Specific Innate Immune Response. Mol Biol Evol 32:1717-1729.
West AP, Shadel GS, Ghosh S. 2011. Mitochondria in innate immune responses. Nat Rev Immunol 11:389-402.
Williams AH, Sharma M, Thatcher LF, Azam S, Hane JK, Sperschneider J, Kidd BN, Anderson JP, Ghosh R, Garg G, et al. 2016. Comparative genomics and prediction of conditionally dispensable sequences in legume-infecting Fusarium oxysporum formae speciales facilitates identification of candidate effectors. BMC Genomics 17:191.
Williams GC. 1957. Pleiotropy, Natural Selection, and the Evolution of Senescence Evolution 11:398-411.
Williams SA, Shattuck MR. 2015. Ecology, longevity and naked mole-rats: confounding effects of sociality? Proc Biol Sci 282.
Wilmoth JR. 2000. Demography of longevity: past, present, and future trends. Exp Gerontol 35:1111-1129.
Wilmoth JR, Deegan LJ, Lundstrom H, Horiuchi S. 2000. Increase of maximum life-span in Sweden, 1861-1999. Science 289:2366-2368.
Wu Q, Zheng P, Hu Y, Wei F. 2014. Genome-scale analysis of demographic history and adaptive selection. Protein Cell 5:99-112.
Xia X. 1996. Maximizing transcription efficiency causes codon usage bias. Genetics 144:1309-1320.
Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. Mol Biol Evol 15:568-573.
Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24:1586-1591.
Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput Appl Biosci 13:555-556.
Yang Z. 2005. The power of phylogenetic comparison in revealing protein function. Proc Natl Acad Sci U S A 102:3179-3180.
Yang Z, dos Reis M. 2011. Statistical properties of the branch-site test of positive selection. Mol Biol Evol 28:1217-1228.
Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. Genetics 141:1641-1650.
Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19:908-917.
Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol 22:1107-1118.
Yokoyama S. 2013. Synthetic biology of phenotypic adaptation in vertebrates: the next frontier. Mol Biol Evol 30:1495-1499.
Yokoyama S, Tada T, Zhang H, Britt L. 2008. Elucidation of phenotypic adaptations: Molecular analyses of dim-light vision proteins in vertebrates. Proc Natl Acad Sci U S A 105:13480-13485.
Zhang C, Cuervo AM. 2008. Restoration of chaperone-mediated autophagy in aging liver improves cellular maintenance and hepatic function. Nat Med 14:959-965.
Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22:2472-2479.
Zhou T, Ko EA, Gu W, Lim I, Bang H, Ko JH. 2012. Non-silent story on synonymous sites in voltage-gated ion channel genes. PLoS One 7:e48541.

# Danksagung

Zunächst will ich allen Koautoren danken. Ohne sie wären die wissenschaftlichen Publikationen, die dieser kumulativen Dissertation zu Grunde liegen, nicht möglich gewesen. In diesem Zusammenhang möchte ich von den Mitgliedern meiner Arbeitsgruppe „Genomanalyse/Platzer-Labor“ insbesondere die folgenden erwähnen: Dr. Marco Groth, der – von den verwendeten, öffentlichen Daten abgesehen – alle Proben sequenziert hat, die in diese Arbeit eingeflossen sind; Martin Bens, der diese Sequenzdaten für meine nachfolgenden Analysen zur positiven Selektion entscheidend weiterverarbeitet hat und mit dem ich in der täglichen Zusammenarbeit der letzten vier Jahren zahllose, kleinere und größere, wissenschaftliche und sonstige Doktorandenprobleme gelöst oder zumindest erörtert habe; Karol Szafranski, der als Mentor und Ausbilder sein Bestes gegeben sowie großen Anteil daran hat, dass meine Doktorandenstelle ins Leben gerufen wurde. Für Letzteres danke ich auch dem Kooperationspartner Philip Dammann, der u.a. maßgeblich in die Haltung und Beprobung verschiedener Nagetierspezies involviert war, die in dieser Arbeit eine Rolle spielen. Die ebenso ideenreiche wie zielorientierte Arbeit des Kooperationspartners Alessandro Cellerino bei mehreren der hier aufgeführten Manuskripte war nicht nur entscheidend für deren Publikation, sondern auch lehrreich und inspirierend für mich als Nachwuchswissenschaftler.

Mein Dank gilt ebenfalls allen noch nicht namentlich erwähnten Mitgliedern der Arbeitsgruppe für die kollegiale Zusammenarbeit in der gemeinsamen Zeit: PD Dr. Klaus Huse, Niels Jahn, Dr. Kathrin Reichwald, Maja Dziegelewska, Silke Förste, Dr. Marius Felder, Dr. Stefan Taudien, Susanne Fabisch, Cornelia Luge, Dr. Philipp Koch, Beate Szafranski und Patricia Möckel. Ganz besonders möchte ich unserem Administrator Bernd Senf sowie Ivonne Görlich, die für die Präparierung der erwähnten Proben im Labor verantwortlich zeichnet, danken.

Ich erkenne ferner die Zeit und die Mühe an, die mein Promotions-Komitee bestehend aus PD Dr. Matthias Platzer, Prof. Dr. Rainer König und Prof. Dr. Christoph Englert investiert hat.

Für die vielen nächtlichen Stunden, die er mit der sprachlichen Kontrolle dieser Arbeit verbracht hat, möchte ich mich bei meinem Vater, Toralf Pfannenstill, bedanken.

Abschließend danke ich meinem Betreuer PD Dr. Matthias Platzer für seine Rolle als wichtigster Ideen- und Ratgeber, Lektor der von mir verfassten Texte sowie vor allem als der maßgebliche Konstrukteur aller beschriebenen Kooperationen innerhalb und außerhalb der Arbeitsgruppe, auf denen diese Arbeit beruht.

# Lebenslauf

| | |
|---|---|
| Name | Arne Sahm |
| Geburtsdatum | 1.10.1989 |
| Geburtsort | Königs Wusterhausen (Brandenburg) |
| Staatsbürgerschaft | Deutsch |

| Ausbildungsniveau | Zeitraum | Einrichtung | Abschluss |
|---|---|---|---|
| **Grundschule** | 1996 – 2002 | Humboldt-Grundschule Eichwalde (Brandenburg) | – |
| **Gymnasium** | 2002 – 2009 | Humboldt-Gymnasium Eichwalde (Brandenburg) | Abitur (1,4) |
| **Universität (Bachelor)** | 2009 - 2012 | Universität Bielefeld (Nordrhein-Westfalen) | Bioinformatik und Genomforschung Bachelor of Science (1,6) |
| **Universität (Master)** | 2012 - 2013 | Universität Bielefeld (Nordrhein-Westfalen) | Bioinformatik und Genomforschung Master of Science (1,3) |
| **Promotiom** | 2013 - heute | Leibniz-Institut für Alternsforschung - Fritz-Lipmann-Institut e.V. (Thüringen) | Dr. rer. nat. (*summa cum laude*) |

# Ehrenwörtliche Erklärung

Hiermit erkläre ich, dass mir die geltende Promotionsordnung der Biologisch-Pharmazeutischen Fakultät der Universität Jena bekannt ist. Ich erkläre ferner, dass ich keine Textabschnitte eines Dritten oder eigener Prüfungsarbeiten ohne Kennzeichnung übernommen und alle von mir benutzten Hilfsmittel und Quellen angegeben habe. Bei der Erstellung der Dissertation wurde ich ausschließlich von den Koautoren und den in den jeweiligen Danksagungen erwähnten Personen der eingebundenen Manuskripte unterstützt. Weder habe ich einen Promotionsberater in Anspruch genommen, noch Dritten mittelbar oder unmittelbar geldwerte Leistungen für Arbeiten im Zusammenhang mit dem Inhalt der vorgelegten Dissertation zukommen lassen. Die Dissertation wurde zuvor für keine andere staatliche oder wissenschaftliche Prüfung eingereicht – auch nicht in ähnlicher Form.

Jena, 30.06.2017

_____

Arne Sahm